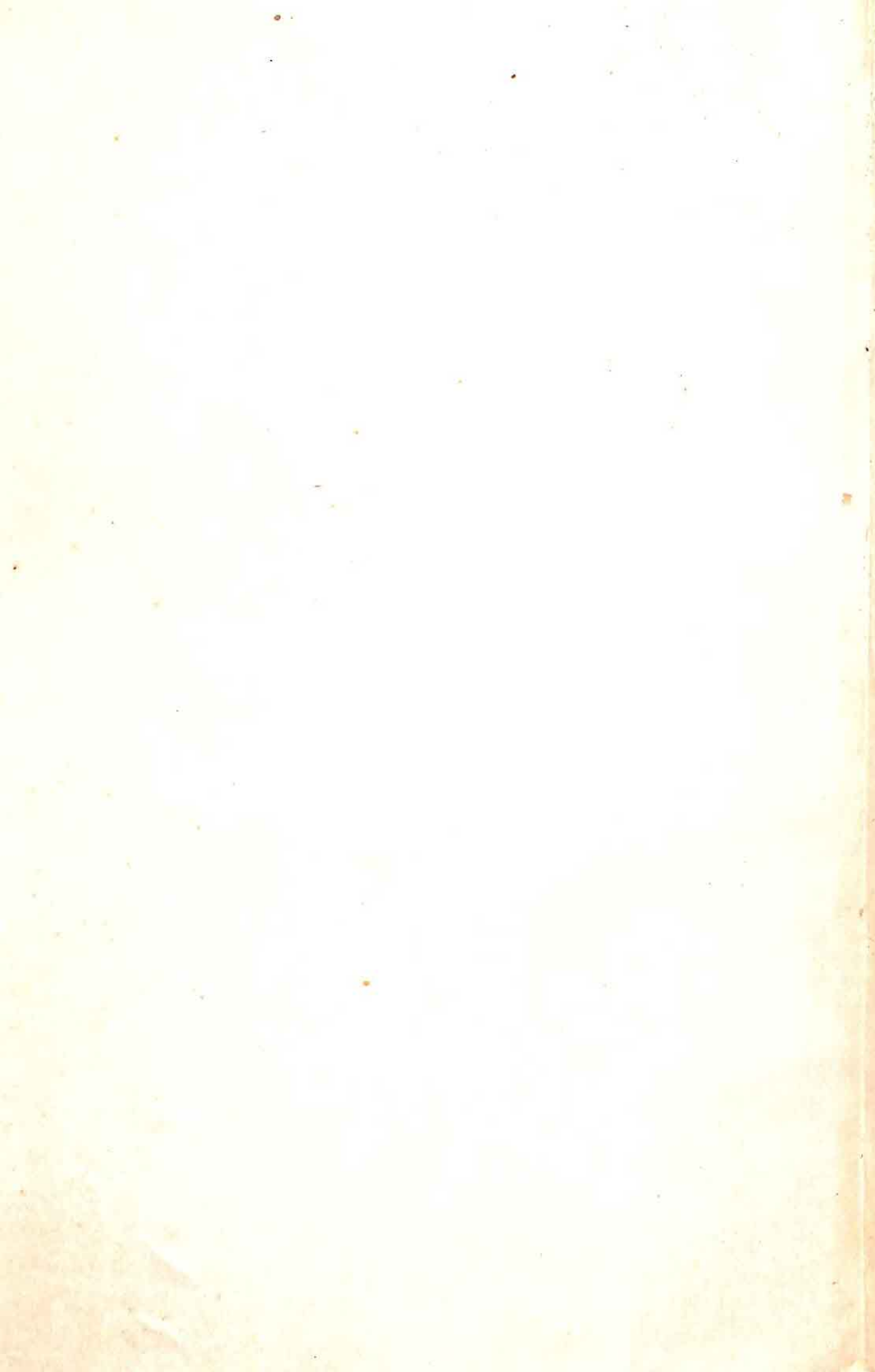


GROUNDWORK OF MATHEMATICAL PROBABILITY AND STATISTICS

AMRITAVA GUPTA

ACADEMIC PUBLISHERS • CALCUTTA • NEW DELHI



3851
10.3.87

GROUNDWORK OF
MATHEMATICAL PROBABILITY AND STATISTICS

Debajyoti Das

Statistics in Biology and Psychology

1c

GROUNDWORK OF MATHEMATICAL PROBABILITY AND STATISTICS

AMRITAVA GUPTA

Reader in Applied Mathematics, University of Calcutta



**ACADEMIC PUBLISHERS
CALCUTTA • NEW DELHI**

© Reserved by the author

First Edition August 1962

Second Edition September 1971

Third Edition April 1983

Price rupees forty only

T. West Bengal

10-3-87

3851

519.9

6UP

The paper used for printing this book has been made available at concessional rates by the Government of India.

ACADEMIC PUBLISHERS

5A, Bhawani Dutta Lane, Calcutta-700073

and also at

SHANTIMOHUN HOUSE,

I-1/16, Ansari Road, New Delhi-110002

To My Students
who gave me the opportunity of learning the subject

PREFACE TO THE THIRD EDITION

After being out of print for quite some time, here is an enlarged third edition. The enlargement in the main consists in providing partial solution or broad hints to all relatively difficult problems given as exercises along with their answers. This would presumably be helpful to the young learners. In Chapter 4 a section has been added on Markov Chain which is interesting theoretically as well as in applications.

As in the second edition of the book, use has been made of the simple notion of monotonic sequences of sets, with the help of which the basic properties of the probability function and the distribution function have been rigorously established. In fact, it is observed that if we are prepared to ignore the rather delicate question of measurability, we can develop without much effort and complexity a mathematical theory of probability which is logically quite accurate and fairly complete.

I wholeheartedly thank the Academic Publishers for their keen interest and sincere co-operation in the publication of the present edition of the book.

AMRITAVA GUPTA

Calcutta, April, 1983

PREFACE TO THE FIRST EDITION

The present-day theories of probability and statistics have been placed on a sound logical basis, but a rigorous exposition of the same requires higher mathematical tools like the concepts of measure and integration which are used in almost all the authoritative works on the subject. On the other hand, the elementary text-books, perhaps with a view to avoiding difficult mathematics, do not pay adequate attention to the logical content and seek only to give a compilation of the various working formulae derived mainly from intuitive considerations. As such a beginner who is not equipped enough to read the authoritative works has to depend on these elementary text books, and a searching mind often finds itself uncomfortable against the vague and loose concepts presented therein. Accordingly, in this book I have set myself to the task of giving a logically satisfactory and complete account of the general principles of the subject as much as possible within the reach of relatively simple mathematical tools. A simple knowledge of elementary analysis is a sufficient prerequisite for reading this book ; a few other mathematical tools necessary, i.e. simple notions of sets, step functions etc. have been developed in the course of the text.

I have started with the idea of random experiments and event spaces and subsequently developed an axiomatic theory of probability based on a set of simple axioms, particular emphasis being laid on the fundamental concepts and the logical coherence of the development. The problem of regression has been treated in some details, and certain departures from the traditional modes of treatment will be noticed in this connection. It was a pleasant surprise to find that correlation ratio can be defined as a correlation coefficient, from which many of its important properties immediately follow. The fundamental limit theorems like the central limit theorem, the continuity theorem for characteristic functions etc. have been assumed without proof, and only the simple consequences discussed.

In statistics, although our major pre-occupation is with the mathematical aspects of the subject, some amount of descriptive elements and computational procedures have also been incorporated. In testing of hypotheses, the general Neyman-Pearson theory of best critical region and the method of likelihood ratio testing have been propounded, and all the important tests are deduced directly from these general principles. I have added a chapter on the theory of errors, which is of interest to many, treated in terms of modern statistical concepts and terminology. An attempt has also been made to give a unified version of the principle of least squares which does not usually appear in exactly the same form in different contexts like regression, theory of errors etc.

I have made use of almost all the books I could lay hands on. The influences of Cramer, Feller, Kendall and Wilks, in particular, will be clearly discernible in the text. A bibliography appears at the end of the book.

I am indebted to Sir Ronald A. Fisher, F. R. S., Cambridge and to Dr. Frank Yates, F. R. S., Rothamsted, also to Messrs. Oliver & Boyd Ltd., Edinburgh, for permission to reprint Tables No. II, III, IV, V from their book 'Statistical Tables for Biological, Agricultural and Medical Research'. I acknowledge debt to Messrs. R. P. Sarker, A. K. Sarkar and D. Sen for supplying me with many interesting statistical data. I owe particular gratitude to my revered teacher and colleague Dr. B. S. Ray whose constant inspiration was a guiding force throughout the preparation of the book. I heartily thank Messrs. S. Ray Chaudhury, J. Paul and P. K. Chatterjee for their assistance in proof-reading and the publishers and the printers for their sincere co-operation. Lastly, a great share of my thanks giving is kept in store for those who will in future help me by offering criticisms and suggestions for improvement.

AMRITAVA GUPTA

Calcutta, August 1962

CONTENTS

MATHEMATICAL PROBABILITY

CHAPTER 1. Event Spaces	1
1.1 Random experiments or observation—1.2 Events-simple and compound—1.3 Mathematical tools: preliminary notions of sets—1.4 The event space—1.5 Exercises	
CHAPTER 2. Historical Background	12
2.1 Introduction—2.2 The classical definition—2.3 Statistical regularity and the frequency definition of probability	
CHAPTER 3. Fundamental Axioms	21
3.1 Axioms of mathematical probability—3.2 Conditional probability—3.3 Stochastic independence—3.4 Exercises	
CHAPTER 4. Compound Experiments	46
4.1 Cartesian product of sets—4.2 Joint independent experiments—4.3 Repeated independent trials—4.4 Bernoulli trials—4.5 Poisson trials—4.6 Multinomial law—4.7 Infinite sequence of Bernoulli trials—4.8 Markov chains—4.9 Exercises	
CHAPTER 5. Probability Distributions	66
5.1 Mathematical tools: functions on sets—5.2 Random variables—5.3 Distribution function—5.4 Mathematical tools: step-function—5.5 Discrete distributions—5.6 Important discrete distributions—5.7 Continuous distributions—5.8 Important continuous distributions—5.9 Transformation of random variables—5.10 Exercises	
CHAPTER 6. Two-dimensional Distributions	94
6.1 Distribution function in two dimensions—6.2 Discrete distributions—6.3 Continuous distributions—6.4 Important two-dimensional or bivariate continuous distributions—6.5 Conditional distributions—6.6 Transformation of random variables in two dimensions—6.7 Extensions to many dimensions. Mutual independence—6.8 Exercises	
CHAPTER 7. Mathematical Expectations I	127
7.1 Mathematical expectation or mean value—7.2 Mean—7.3 Moments—7.4 Variance—7.5 Third central moment—7.6 Fourth central moment—7.7 Moment generating function—7.8 Charac-	

teristic function—7.9	Semi-invariants or cumulants—7.10	
Median—7.11	Mode—7.12	Quantiles—7.13
Some remarks—7.14	Exercises	
CHAPTER 8. Mathematical Expectations II		152
A. Two-dimensional Case—8.1	Expectation for a bivariate distribution—8.2	Moments—8.3
Covariance, correlation coefficient—8.4	Characteristic function—8.5	Some extensions to n -dimensions
B. Independent Random Variables—8.6	Multiplication rule for expectations—8.7	Moments—8.8
Characteristic function—8.9	Another discussion on Bernoulli trials	
C. Conditional Expectations and Regression—8.10	Conditional expectation—8.11	Regression curves—8.12
Least square regression curves—8.13	Regression lines—8.14	Parabolic curve fitting—8.15
Correlation ratio—8.16	Exercises	
CHAPTER 9. Special Distributions		188
9.1 χ^2 -distribution—9.2	t -distribution—9.3	F -distribution—9.4
Exercises		
CHAPTER 10. Convergence 'in Probability'		197
10.1 Tchebycheff's inequality—10.2	Convergence 'in probability'—10.3	Exercises
CHAPTER 11. Limit Theorems		204
11.1 Normal approximation to the binomial distribution—11.2	Fundamental limit theorems—11.3	Exercises
MATHEMATICAL STATISTICS		
CHAPTER 12. Random Samples		215
12.1 Populations and samples—12.2	Distribution of the sample—12.3	Tables and graphical representations—12.4
Sample characteristics—12.5	Computation of sample characteristics—12.6	Exercises
CHAPTER 13. Sampling Distributions		230
13.1 Sampling distributions of 'statistic's—13.2	Estimates-consistent and unbiased—13.3	Important sampling distributions—13.4
Normal population—13.5	Exercises	
CHAPTER 14. Estimation of Parameters		241
14.1 Method of maximum likelihood—14.2	Applications to different populations—14.3	Interval estimation—14.4
Method for finding confidence intervals—14.5	Applications to normal (m, σ) population—14.6	Approximate confidence intervals—14.7
Exercises		

CHAPTER 15. Bivariate Samples	257
15.1 Sample from a bivariate population—15.2 Practical computation—15.3 Least square curve fitting—15.4 Maximum likelihood estimation—15.5 Exercises	
CHAPTER 16. Testing of Hypotheses I	271
16.1 Statistical hypotheses-simple and composite—16.2 General form of a test. Best critical region—16.3 Best critical regions for simple hypotheses—16.4 Applications to normal (m, σ) population—16.5 Likelihood ratio testing—16.6 Normal (m, σ) population—16.7 Comparison of normal populations—16.8 Bivariate normal population—16.9 Exercises	
CHAPTER 17. Testing of Hypotheses II	302
17.1 Binomial (n, p) population—17.2 Comparison of binomial populations—17.3 Poisson- μ population—17.4 Multinomial distribution—17.5 Multinomial population—17.6 χ^2 -test of goodness of fit—17.7 Exercises	
CHAPTER 18. Theory of Errors	316
18.1 Introduction—18.2 The normal law—18.3 Some definitions—18.4 Estimation—18.5 Weighted measurements—18.6 Indirect observations—18.7 Exercises	

MATHEMATICAL PROBABILITY

EVENT SPACES

1.1 RANDOM EXPERIMENTS OR OBSERVATIONS

The word 'probability' figures very often in our everyday speech and in a wide variety of contexts; for example, 'probably it will rain to-morrow', 'probably he is an honest man', 'the probability that there will be a bumper crop in the next season is very small', 'what is the probability of a double six in a throw of a pair of dice?' and so on. Any attempt towards a theory of probability naturally begins with the question as to the probability of what we are interested in. The immediate answer is obviously—*events*. This, however, only introduces a general name which is in no way self-explanatory. Our first task then will be to make precise the meaning of the term 'event' and the proper context in which it will be used in our mathematical theory. For this, we come to the idea of what are called *random experiments* or *observations*.

Let us take the case of tossing a coin. We know that there are two possible outcomes—'head' and 'tail', and that it is impossible to predict if the result of a toss will be a 'head' or 'tail'. Consider a similar experiment of rolling a die from a box; there are only six possible results, viz. the faces marked 1, 2, ..., 6, but here also the result of a particular throw is completely unpredictable. Or suppose we are concerned with the measurement of a physical quantity by means of a precision instrument. Students of physics know that the result of a measurement does not exactly give the true value of the quantity but a value close to it due to what are called *experimental errors*. If repeated observations are taken, the measured values are not again the same but fluctuate in an unpredictable manner. Here we can take, at least for theoretical considerations, that the possible results comprise all the real numbers, but the number given by a single measurement cannot be exactly predicted. In our mathematical theory, we shall only consider the class of those experiments or

observations, for which we know a priori the set of all different possible results or outcomes, and which are such that it is impossible to predict which one of this set will occur at any particular performance of the experiment. Such experiments are called *random experiments*, the word 'random' pertaining to the above-mentioned lack of predictability. As such, if a random experiment is repeated under identical conditions, the results will vary at random.

So far we have said nothing about the reasons for this randomness. The reasons are, however, manifold and not always clearly understood and, for that matter, not also essential for our theory of probability. Our theory, in fact, starts with accepting this idea of randomness and need not explain it. Still, in order to get a deeper glimpse into the situation, let us consider the process of, say, throw with a die. The die is shaken well in a box and thrown on a table, and suppose that the first result is 'six'. If now the die is thrown again under identical conditions, the result will, however, be not necessarily 'six', but may be any one of the six faces. This may seem somewhat paradoxical if we are pondering that the mechanical behaviour of the die should be uniquely determined by the initial conditions of throwing and, of course, the laws of mechanics, and, as such, if the die is thrown under identical conditions, the results must also be the same. The explanation, however, lies in the fact that, although throwing a die looks a simple affair, it is a very complicated mechanical process, and it is practically impossible to create *exactly* identical initial conditions of throw. These conditions vary, however subtly, at random, and this produces the unpredictable variability of results in a sequence of repetitions of the experiment. Thus what we mean by creating identical conditions is only keeping the relevant conditions of the experiment *as uniform as possible*, and let us bear in mind that the phrase *identical conditions* will henceforth be used only in this approximate sense.

Consider now a slightly different case. Let two equally massive particles be moving in a given field of force, being let go from the same initial position and with the same initial velocity. If it is known that the initial conditions are really identical, and if it is observed that at a subsequent instant the particles occupy different positions, we are faced with yet another type of and a more difficult logical

problem. It will be interesting to mention that such a problem is not a hypothetical one, but was actually observed by the physicists in a very small scale, viz. the atomic scale. Physicists were, however, divided in their approach towards the explanation of such a phenomenon. One school of thought took up a radically new outlook and proclaimed that the laws of Nature are possibly not exactly fixed and can make room for small random fluctuations which make themselves felt in a small scale. This philosophy goes by the name of *Principle of Indeterminacy*. The orthodox school, however, continued to believe that Nature must be guided by perfectly deterministic laws, but it might be that the classical laws break down in a small scale and need be replaced by possibly more complicated laws necessitating the use of more complicated initial conditions.

To sum up, for reasons known or unknown there must be some *intrinsic variability* in the process of our experiments which would make them random. If, by increasing the perfection of the experimental process or otherwise, it is possible to get rid of this variability so that particular results become predictable, the experiment ceases to be random and is naturally pushed out of the realm of our probability theory.

1.2 EVENTS—SIMPLE AND COMPOUND

The outcomes or results of a random experiment will be called *events connected with the experiment*, e.g. 'head' and 'tail' are results of the random experiment of throwing a coin and hence are events connected with it. We can distinguish between two types of events—simple and compound. To understand this, consider the experiment of rolling a die; 'one', 'two', ... 'six' are certainly events connected with the experiment. Now the result 'six' can also be described under a different title, say, 'even face' or 'multiple of three', only that in the latter cases the result 'six' is not uniquely specified. Thus 'even face' and 'multiple of three' are also events. But the event 'even face' not only occurs when the result is 'six' but also when the result is 'two' or 'four', and we say that the event 'even face' can be *decomposed* into the events 'two', 'four' and 'six'. Similarly, the event 'multiple of three' can be decomposed into the events 'three' and 'six'. The events 'one', 'two' etc. cannot, however, be further

decomposed. Events which cannot be further decomposed are called *simple events*, and *compound events* are those which can be decomposed into simple events.

An event which is sure to occur at every performance of the experiment is called a *certain event*. For example, 'one or two or three...or six' is a certain event in connection with throw of a die. We may also think of events which are logically impossible, i.e. which cannot occur at any performance of the experiment. Such events will be called *impossible events*. The event 'seven' can never be the result of throwing a die and is thus an impossible event. Clearly, *a certain event can be decomposed into all the possible simple events, while an impossible event cannot be decomposed into any one of them.*

If when an event occurs another event invariably occurs, then the former event is said to *imply* the latter event. For example, the event 'two or four' implies the event 'even face' for throwing a die. If an event implies another event, then the simple events into which the first event can be decomposed are also some of the simple events into which the second event can be decomposed. Obviously, *any event implies the certain event.*

Two events are said to be *equivalent* or *identical* if any one of them implies and is implied by the other. Thus the events 'two or four or six' and 'even face' are identical.

An event may be titled 'either even face or multiple of three or both'; this is a compound event which can be decomposed into the four simple events 'two', 'three', 'four' and 'six'.

The events 'even face' and 'multiple of three' occur simultaneously when and only when the result is 'six'. In other words, we may say that the event 'joint occurrence of even face and multiple of three' can be decomposed into the simple event 'six'.

If two events are such that they cannot occur simultaneously, they are said to be *mutually exclusive*, e.g. 'even face' and 'odd face' are mutually exclusive events. We note that *two simple events are always mutually exclusive*, but compound events may or may not be so. Thus the compound events 'even face' and 'multiple of three' are not mutually exclusive.

An event which consists in the negation of another event is called the *complementary event* of the latter event. The complementary event of 'multiple of three' is obviously 'not multiple of three' which can be decomposed into the simple events 'one', 'two', 'four' and 'five'. Note that *the complementary event of a certain event is an impossible event and vice versa*.

For recollection, take another example. Let the experiment consist in drawing a card at random from a well-shuffled pack of playing cards. There are 52 possible simple events. The event 'spade' can be decomposed into 13 simple events, and the event 'queen' into 4 simple events. The event 'king or queen or jack of spades' implies the event 'spade'. The events 'spade' and 'queen' are not mutually exclusive, and the event of their joint occurrence is the simple event 'queen of spades'. The event 'either spade or queen or both' will be decomposed into 16 simple events. 'Eleven of spade' is an impossible event; 'any card' is a certain event. The complementary event of 'spade' is 'club or heart or diamond'.

For a precise mathematical formulation of the concept of events connected with a random experiment, we would require some knowledge of the theory of sets which we now develop.

1.3 MATHEMATICAL TOOLS : PRELIMINARY NOTIONS OF SETS

The aggregate or collection of all possible objects having given properties will be called a *set*. The objects belonging to the set are called *elements* of the set. For example, the set of chairs in a particular room, the set of non-negative integers, the set of real numbers x such that $a < x < b$ etc.

When an element a belongs to a set S , we write in symbols $a \in S$.

If every element of a set A belongs to a set S , then we say that A is *contained in* S , or S *contains* A , or that A is a *subset* of S and write symbolically $A \subseteq S$ or $S \supseteq A$.

If $A \subseteq S$ and $S \subseteq A$, i.e. every element of A belongs to S and every element of S belongs to A , then we say that the sets A and S are *identical* or *equal* and write $A = S$.

If $A \subseteq S$, but $A \neq S$, i.e. every element of A belongs to S but there is at least one element of S which does not belong to A , then A is said to be a *proper subset* of S and written as $A \subset S$.

A *null* or an *empty* set is one which does not contain any element at all and will be denoted by O .

We note the following :

1. Every set is a subset of itself.
2. An empty set is a subset of every set.
3. A set containing only one element is conceptually distinct from the element itself but will be represented by the same symbol for the sake of convenience.

Often we are concerned with the study of various subsets of a given set S . In such cases, it is customary to use a geometric terminology, in which the elements of S are called *points* and the set S is called a *space*. Let S be a given space, and $A, B, C, \dots; A_1, A_2, A_3, \dots$ denote subsets of S .

The *sum* or *union* of two sets A and B is denoted by $A+B$ or $A \cup B$ and is defined to be the set of all elements belonging to either A or B or both. Note that $A+B$ is also a subset of S .

The *product* or *intersection* of two sets A and B is denoted by AB or $A \cap B$ and is defined to be the set of all elements belonging to both A and B . Then $AB \subseteq S$. If $AB=O$, the sets A and B are said to be *disjoint*.

It is easily seen that the following laws hold for the above-defined *addition* and *multiplication* of sets.

- | | | |
|-------|------------------------------------|----------------------|
| (i) | $(A+B)+C=A+(B+C)$
$(AB)C=A(BC)$ | } (associative laws) |
| (ii) | $A+B=B+A$
$AB=BA$ | |
| (iii) | $A(B+C)=AB+AC$ | (distributive law) |

By virtue of (i) and (ii) we can write without ambiguity

$$A_1 + A_2 + \dots + A_n \text{ and } A_1 A_2 \dots A_n$$

where the order of terms or factors is arbitrary. Thus the sum $A_1 + A_2 + \dots + A_n$ is the set of all elements belonging to at least one of

the sets A_1, A_2, \dots, A_n , and the product $A_1 A_2 \dots A_n$ is the set of all elements belonging to each one of the sets A_1, A_2, \dots, A_n .

Let $B \subseteq A$. The *difference* $A - B$ is defined to be the set of elements of A which do not belong to B . In particular, the set $S - A$ will be called the *complement of A in S* and will be denoted by \bar{A} , i.e. $\bar{A} = S - A$. It follows obviously

$$A + \bar{A} = S, \quad A\bar{A} = O \quad \text{and} \quad \bar{\bar{A}} = A \quad (1.3.1)$$

For any two sets A and B we may write

$$A + B = (A - AB) + AB + (B - AB) \quad (1.3.2)$$

where the sets $A - AB$, AB and $B - AB$ are pairwise disjoint. This formula can be verified from the following diagram (Fig. 1), in which the region bounded by the bold line represents the set $A + B$, the shaded region represents AB , and the unshaded parts of A and B the sets $A - AB$ and $B - AB$ respectively.

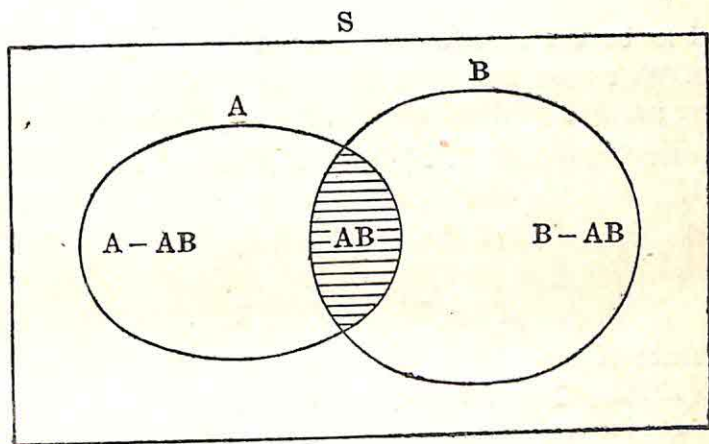


Fig. 1

Also note the following interesting results :

$$\begin{aligned} A + A &= A, & AA &= A \\ S + A &= S, & SA &= A \\ \bar{S} &= O, & \bar{O} &= S \end{aligned} \quad (1.3.3)$$

We may easily prove that

$$\begin{aligned} \overline{(A + B)} &= \bar{A} \bar{B} \\ \overline{AB} &= \bar{A} + \bar{B} \end{aligned} \quad (1.3.4)$$

Generalising to n sets A_1, A_2, \dots, A_n we have, if

$$X = A_1 + A_2 + \dots + A_n$$

$$Y = A_1 A_2 \dots A_n$$

then

$$\overline{X} = \overline{A_1} \overline{A_2} \dots \overline{A_n} \quad (1.3.5)$$

$$\overline{Y} = \overline{A_1} + \overline{A_2} + \dots + \overline{A_n}$$

Sequences of sets

Consider now an infinite sequence $\{A_n\}$ of sets $A_n \subseteq S$ ($n = 1, 2, \dots$).

The sum or union

$$A_1 + A_2 + \dots + A_n + \dots = \sum_{n=1}^{\infty} A_n$$

is defined to be the set of all elements which belong to A_n for at least one value of n , and the product or intersection

$$A_1 A_2 \dots A_n \dots = \prod_{n=1}^{\infty} A_n$$

is defined to be the set of all elements which belong to A_n for every value of n . We remark that these definitions of infinite sum and infinite product are purely logical and are free from any limiting process.

The generalisations of (1.3.5) for a sequence of sets, $\{A_n\}$ will be formal. If

$$X = \sum A_n, \quad Y = \prod A_n$$

then

$$\overline{X} = \prod \overline{A_n}, \quad \overline{Y} = \sum \overline{A_n} \quad (1.3.6)$$

A sequence of sets $\{A_n\}$ is said to be *monotonic non-decreasing* or *expanding* if $A_{n+1} \supseteq A_n$ for every value of n , and $\{A_n\}$ is said to be *monotonic non-increasing* or *contracting* if $A_{n+1} \subseteq A_n$ for every n .

Let $\{A_n\}$ be an expanding sequence of sets. Then $\lim A_n$ as $n \rightarrow \infty$ is defined by

$$\lim A_n = \sum A_n$$

If $\{A_n\}$ is a contracting sequence of sets, then $\lim A_n$ is defined by

$$\lim A_n = \prod A_n$$

It may be easily seen that if $\{A_n\}$ is an expanding (or contracting) sequence, then $\{\overline{A_n}\}$ is a contracting (or expanding) sequence, and it follows from (1.3.6) that if $\{A_n\}$ is either expanding or contracting

$$\overline{\lim A_n} = \lim \overline{A_n} \quad (1.3.7)$$

1.4 THE EVENT SPACE

Let the random experiment be denoted by E .

The simple events connected with E will be called *event points* or simply *points*, and the set of all possible event points the *event space* S of E . Then the experiment E must be such that its event space S is completely known.

Any subset A of S will be called an *event connected with E* , i.e. an event is an aggregate of some of the event points.

The entire space is the *certain event*, and the empty subset O is the *impossible event*.

We say that an event A *implies* another event B if the set A is contained in the set B , i.e. $A \subseteq B$. The events A and B are said to be *equivalent* or *identical* if the sets A and B are identical.

For any two events A, B , the event '*either A or B or both*' is defined to be the set $A + B$, and the event ' *A and B occurring simultaneously*' to be the set AB . The events A and B are said to be *mutually exclusive* if the sets A, B are disjoint, i.e. $AB = O$.

Let A_1, A_2, \dots be any finite or infinite sequence of events. Then the sum $A_1 + A_2 + \dots$ will be called the event of *occurrence of at least one of the events A_1, A_2, \dots* and the product $A_1 A_2 \dots$ the event of *simultaneous occurrence of all the events A_1, A_2, \dots* .

\bar{A} , the complement of A in S will be naturally called the *complementary event of A* . Since $\bar{\bar{A}} = A$, it follows that *the complementary event of a complementary event is the event itself*. Also, since $A\bar{A} = O$ and $A + \bar{A} = S$, *two complementary events are mutually exclusive, and their sum is the certain event*. The formulae $\bar{S} = O$, $\bar{O} = S$ immediately show that *the impossible and the certain events are complementary events*.

With obvious meanings we shall speak of *expanding* and *contracting sequences of events* and their *limiting events*.

Examples

1. Let E denote the experiment of tossing a coin three times in succession. A typical event point is, say, 'head, head, tail' which may

be denoted by the symbol (H, H, T) . The event space S consists of 8 points U_1, U_2, \dots, U_8 given by

$$\begin{aligned} U_1 &= (H, H, H), & U_2 &= (H, H, T), & U_3 &= (H, T, H) \\ U_4 &= (T, H, H), & U_5 &= (T, T, H), & U_6 &= (T, H, T) \\ U_7 &= (H, T, T), & U_8 &= (T, T, T) \end{aligned}$$

and we write

$$S = U_1 + U_2 + \dots + U_8$$

Let A denote the event 'two heads'. Then A contains the 3 points U_2, U_3, U_4 , i.e. $A = U_2 + U_3 + U_4$.

If B be the event 'head in the first trial', then

$$\begin{aligned} B &= U_1 + U_2 + U_3 + U_7 \\ A + B &= U_1 + U_2 + U_3 + U_4 + U_7 \\ AB &= U_2 + U_3 \\ A - AB &= U_4 \\ B - AB &= U_1 + U_7 \end{aligned}$$

We note that the events AB , $A - AB$ and $B - AB$ are pairwise mutually exclusive, and formula (1.3.2) may be easily verified.

The event 'no head or all tails' is obviously the event point U_8 , so that the event 'at least one head' is the complementary event

$$\bar{U}_8 = S - U_8 = U_1 + U_2 + \dots + U_7$$

Remark. It is perhaps clear that, while writing summation with U 's, the symbols do not exactly denote the event points but the sets containing the individual points, so that a sum of U 's is understood in the usual sense of sum of sets. It is again in the latter sense that event points will denote events in our theory.

2. Let E denote the experiment of placing two balls at random into three cells. Now two cases will arise according as the balls are distinguishable among themselves or not. It is, however, assumed that the given cells are distinct.

(a) **DISTINGUISHABLE BALLS.** In this case the balls may be represented by the symbols B_1, B_2 . The event space S will contain the following 9 points :

$$\begin{aligned} U_1 &= (B_1 | B_2 | -), & U_2 &= (B_1 | - | B_2), & U_3 &= (- | B_1 | B_2) \\ U_4 &= (B_2 | B_1 | -), & U_5 &= (B_2 | - | B_1), & U_6 &= (- | B_2 | B_1) \\ U_7 &= (B_1 B_2 | - | -), & U_8 &= (- | B_1 B_2 | -), & U_9 &= (- | - | B_1 B_2) \end{aligned}$$

It A denotes the event 'one ball in the second cell', then

$$A = U_1 + U_3 + U_4 + U_6$$

(b) INDISTINGUISHABLE BALLS. If the balls are indistinguishable, the event points may be obtained by dropping the subscripts of the B 's in case (a). On doing this, we note that the event points U_1, U_4 become identical, U_2, U_5 become identical and U_3, U_6 become identical giving only 6 points in the new event space S' , viz.

$$\begin{aligned} U_1' &= (B|B|-), & U_2' &= (B|-|B), & U_3' &= (-|B|B) \\ U_4' &= (BB|-|-), & U_5' &= (-|BB|-), & U_6' &= (-|-|BB) \end{aligned}$$

Remark. We remark once for all that in all such problems the balls will be treated as distinguishable unless stated to the contrary.

3. Let E consist in counting the number of telephone calls on a given trunkline during a fixed interval of time. The possible counts are 0, 1, 2, ..., and there is no upper limit. Hence the event space S consists of the set of all non-negative integers.

4. Let E consist in measuring the length of a rod by a precision instrument. If we assume, for theoretical idealisation, that a measurement may yield any real number, then the event space S will be the set of all real numbers.

5. If E consists in observing the sex of a new-born baby, the event space S contains only two points 'boy' and 'girl'.

1.5 EXERCISES

1. Prove the formula (1.3.4).
 2. Prove the formula (1.3.5).
 3. Show that if the sequence of sets $\{A_n\}$ is expanding, then $\{\overline{A_n}\}$ is a contracting sequence.

4. Let A_n denote the interval $-\infty < x \leq n$ ($n=1, 2, \dots$). Prove that $\{A_n\}$ is an expanding sequence of sets, and

$$\lim A_n = (-\infty, \infty)$$

5. If A_n denotes the interval $-\infty < x \leq -n$ ($n=1, 2, \dots$), then show that $\{A_n\}$ is contracting, and

$$\lim A_n = \emptyset$$

6. If A_n denotes the half-open interval $a - \frac{1}{n} < x \leq a$ ($n=1, 2, \dots$), then show that $\{A_n\}$ is contracting, and $\lim A_n$ is the set containing a only.

7. If A_n denotes the half-open interval $a < x \leq a + \frac{1}{n}$ ($n=1, 2, \dots$), then prove the $\{A_n\}$ is contracting, and $\lim A_n$ is the empty set.

HISTORICAL BACKGROUND

2.1 INTRODUCTION

The history of probability is a very fascinating topic, and the many interesting stories woven round it are largely well-known. We shall, in this chapter, only trace the mathematical development of the concept of probability, which will form the necessary background for the inception of the present-day axiomatic theory. The theory of probability had a humble beginning in the games of chance connected with gambling in France in the 17th century, and since then it has passed through many phases of metamorphoses and has finally emerged as a beautiful and sophisticated branch of mathematics. Towards the beginning of the 19th century, Laplace put forward a formal definition of probability which goes by the name of the *classical definition*. The classical theory thrived mainly on the diverse problems of games of chance and very well served the popular needs. It was, however, subsequently found out that the classical theory suffers from an intrinsic logical weakness and must be placed on a more sound and an entirely new basis in order to meet the requirements of the expanding fields of its application, viz. *statistics, economics, insurance, biology, physics* etc. A new point of view was explored by von Mises only in the 1920's, who gave a new definition of probability which we shall call the *frequency definition*. The frequency definition, which involves a limiting process, although vastly strengthened the logical frame of the theory, again showed signs of mathematical inelegance and operational inconvenience. In 1933, these defects were ultimately got rid of by Kolmogoroff in an *axiomatic* theory of probability. In this book we shall read a simplified version of this axiomatic theory within the scope of relatively simple mathematical tools. In the rest of this chapter, we shall discuss in outline the classical and frequency definitions of probability together with the criticisms levelled against them. This will enable us to realise the necessity and inevitability of the axioms of

the modern theory, which may otherwise seem somewhat strange and arbitrary.

Consider a simple game of chance. A coin will be tossed ; if the result is 'head', I win and if it is 'tail', I lose. What is the chance of my winning ? Any layman, we believe, will answer immediately—it is 50%, and if he is a little more careful, he will add—provided the coin is a true or symmetrical one. Take the game of throwing a die. If it is 'a multiple of three', I win, otherwise I lose. In this game, what should be the correct ratio of betting for honest gambling ? The layman's answer will undoubtedly be—1 : 2 in my favour assuming, however, that the die is symmetrical about all the six faces. In a precise language, we say that the probability of a 'head' in a toss is $\frac{1}{2}$, and that of the event 'multiple of three' connected with throwing a die is $\frac{1}{3}$ in case we make the convention of expressing probabilities as fractions. But if the layman is asked how did he obtain the numbers $\frac{1}{2}$ and $\frac{1}{3}$ in the above games, he will possibly be at a loss to give a proper explanation and will simply say—from *intuition* and *experience*.

2.2 THE CLASSICAL DEFINITION

The classical definition bases itself mainly on intuition. Although intuition is a difficult thing to be analysed, yet, in the above cases, it will be quite easy to trace the law of formation of the numbers $\frac{1}{2}$ and $\frac{1}{3}$. In the random experiment of tossing a coin there are 2 points in the event space, and the event 'head' contains only 1 point. The ratio of the number of points contained in the event 'head' to the total number of points in the event space gives the fraction $\frac{1}{2}$. In the second case, the event space contains 6 points, of which 2 are contained in the event 'multiple of three', and the ratio of 2 to 6 is the required number $\frac{1}{3}$. In both cases, however, we assume that all the points of the event space are mutually symmetrical. Now to explain further how the ratio obtained by the above rule represents the *probability* of an event is perhaps not possible and must be left entirely to intuition. Thus in the classical theory, we have the following definition of probability :

Let E be a random experiment such that its event space S contains a *finite* number, say n , of event points, all of which are known to be *equally likely* or *mutually symmetrical*. If any event

A connected with E contains $m(A)$ of these event points, then the probability of A , denoted by $P(A)$, will be defined by

$$P(A) = \frac{m(A)}{n} \quad (2.2.1)$$

Criticisms of the classical definition

1. The classical definition, although looks simple, has a grave logical flaw. The use of such a definition requires a priori knowledge of the fact that all the event points are equally likely. Let us examine the phrase *equally likely* a little more closely. How to conclude if the event points of a given space are equally likely? The available argument was that, if the event points are mutually symmetric, they may be taken to be equally likely. But the next question immediately crops up (but ironically it was delayed in history for nearly a century!) —mutually symmetric in what respects? This poses a really difficult problem, and it was found after many serious investigations that it is impossible to set forth definite general criteria for mutual symmetry of the event points, and any such set of criteria includes the tacit assumption that the event points are symmetrical in the sense of probability itself. This amounts to begging the concept of probability before we have defined it and is thus a logical vicious circle. Further, in the absence of definite criteria for mutual symmetry, the only way of decision in a particular problem rests plainly on intuition. Now intuitions of different persons cannot be forced to be unique, and consequently there existed a lot of controversies among mathematicians of the classical school. It was found that, even in slightly complicated problems of games of chance, it becomes really difficult to decide, even in a practical way, if the event points are mutually symmetrical or not, and different mathematicians gave different answers to many such problems.

The above difficulties will become apparent if we consider the simple experiment of throw with a die. Let us try to find the conditions under which the six points of this event space would be mutually symmetric. To start with, we would naturally demand that the die should have a perfectly regular cubical shape and should be made from perfectly homogeneous material. But are these conditions sufficient to ensure mutual symmetry of the event points?

In reply to this, we would possibly add, for fear of incompleteness, that the six faces of the die must be symmetrical with respect to all possible kinetic properties, e.g. the centre of gravity of the cube should coincide with its geometrical centre, the twelve moments of inertia corresponding to the rotations of the cube about its twelve edges must all be equal and so on. Although it is difficult, if not impossible, to count on fingers all the kinetic properties exhaustively, we can still ask the pertinent question if *geometrical* and *kinetic symmetries* are sufficient for the purpose and if the principles of mechanics are the only guiding factors for the result of a throw with a die. It is well-known that the veteran gamblers believe in what is called the *luck factor*, and to them it would seem quite reasonable that the six different numbers inscribed on the six faces of an otherwise symmetrical die may yet produce difference in luck! Humour apart, there is still another serious point to consider. It must not be forgotten that talking about the die only does not tell the whole story about the random experiment, which also includes the process of throwing the die from the box. This, as we have already remarked, is a very intricate and uncertain mechanical process, and, as such, it would be indeed impossible to find the conditions of symmetry of the event points for the process of throwing the die. All these arguments sufficiently convince anyone that it is impossible to find appropriate criteria for mutual symmetry, and if we still want to stick stubbornly to the idea of mutual symmetry, we are ultimately obliged to assume that the event points should be symmetrical from the point of view of probability itself, i.e. the phrase *equally likely* becomes synonymous with *equally probable*. This is a great weakness of the classical definition, and no sound mathematical theory can be hoped to be built on such a weak definition.

2. Moreover, the classical theory has a very narrow compass of applications; it is restricted to a small class of event spaces which contain only a finite number of so-called equally likely event points. With the help of such a theory, it will thus be impossible to treat the cases of unsymmetrical event points, e.g. the case of a loaded die or the sex of a new-born baby in which the two event points 'boy' and 'girl' cannot be assumed to be necessarily equally likely and the cases of infinite number of event points, e.g. measure-

ment of a physical quantity and so on. With the development of the theories of statistics and other prospective fields of application of probability, it was observed that the restricted classical event spaces exist almost nowhere outside the relatively unimportant domain of the games of chance, and the classical theory is utterly powerless to cope with the new requirements.

2.3 STATISTICAL REGULARITY AND THE FREQUENCY DEFINITION OF PROBABILITY

It thus became clear that the classical theory must be abandoned altogether, and the new concept of probability must spring from an entirely new premise. Leaving aside *intuition*, the layman's second guiding factor is *experience*. Let us see what our experience says about random experiments. If a coin is to be tossed once, nothing can be predicted about the result, simply because the experiment is random. But if the coin is tossed a large number of times under identical or uniform conditions, it is very interesting to note that we will be able to tell much about the overall results from experience. For example, if the coin is tossed 200 times we can say that the event point 'head' will occur about 100 times or, in other words, the ratio of the number of times 'head' occurs to the total number of experiments will be approximately $\frac{1}{2}$. If the sequence is made longer, say 2000 times, we can safely predict that the above ratio will be very close to $\frac{1}{2}$ and so on.

In general, we have the following empirical fact. Let a random experiment E be repeated under identical or uniform conditions N times, and if an event A connected with E is found to occur $N(A)$ times, then $N(A)$ will be called the *absolute frequency* or simply the *frequency* of A and the ratio $N(A)/N$ the *relative frequency* or the *frequency ratio* of A and denoted by $f(A)$, i.e.

$$f(A) = \frac{N(A)}{N} \quad (2.3.1)$$

It is observed that as N becomes larger and larger, the frequency ratio $f(A)$ gradually tends to become more or less constant. This phenomenon of stability of frequency ratios for long sequences of repetitions of a random experiment is called *statistical regularity*. This may seem very surprising to be justified logically, in view of the

fact that every repetition of the experiment is independent of all other repetitions of the sequence and is quite open to result in any event point. But this was confirmed by many accurate and laborious experiments, which firmly established statistical regularity as an observational fact.

The new school utilised this phenomenon in formulating a definition of probability. In this theory, we postulate that the sequence $f(A) = N(A)/N$ tends to a definite limit as N tends to infinity, and this limit will be called the probability of the event A to be denoted by $P(A)$, i.e.

$$P(A) = \lim_{N \rightarrow \infty} f(A) \quad (2.3.2)$$

Remarks

1. In this theory we require that the random experiments must be such that they can be repeated an indefinitely large number of times under identical conditions. It may be remarked that this imposes only a mild restriction on the random experiments, which is satisfied in almost all problems of practical importance.

2. This definition is, however, broad enough to include unequally likely as well as infinite number of event points.

3. But the real strength and beauty of the theory lies in the fact that a definite operational meaning has been ascribed to probability, viz. for a long sequence of repetitions of a random experiment, the frequency ratio of a given event will be approximately equal to its probability, i.e. the number of times the event will occur is approximately equal to its probability times the total number of experiments. And this is the sense in which we can make use of our knowledge of probability. In contrast to this, the classical definition only intuitively satisfies our feeling for the word *chance* or *probability*.

Deduction of some important rules

1. For any event A , $0 \leq N(A) \leq N$ or dividing by N , $0 \leq f(A) \leq 1$. In the limit as, $N \rightarrow \infty$

$$0 \leq P(A) \leq 1 \quad (2.3.3)$$

2. The frequency of the certain event, $N(S) = N$ or $f(S) = 1$ so that in the limit

$$P(S) = 1 \quad (2.3.4)$$

The frequency of the impossible event, $N(O) = 0$ or $f(O) = 0$. Hence

$$P(O) = 0 \quad (2.3.5)$$

3. Let A and B be any two mutually exclusive events, i.e. $AB = O$. Clearly

$$N(A + B) = N(A) + N(B)$$

or

$$\frac{N(A + B)}{N} = \frac{N(A)}{N} + \frac{N(B)}{N}$$

or

$$f(A + B) = f(A) + f(B)$$

Making $N \rightarrow \infty$

$$P(A + B) = P(A) + P(B) \quad (2.3.6)$$

If A, B, C be pairwise mutually exclusive, i.e. $AB = O, BC = O, CA = O$, then the events A and $B + C$ are also mutually exclusive, for $A(B + C) = AB + AC = O$, and we have

$$P(A + B + C) = P(A) + P(B + C) = P(A) + P(B) + P(C)$$

In general, if A_1, A_2, \dots, A_n be n pairwise mutually exclusive events, i.e. $A_i A_j = O$ ($i \neq j$; $i, j = 1, 2, \dots, n$), we have the following addition rule :

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n) \quad (2.3.7)$$

Conditional probability

Consider two events A and B . Let us make the hypothesis that the event A has occurred. Then in the sequence of N repetitions of the random experiment E , we have to consider only a *subsequence* of $N(A)$ repetitions in which A has occurred, and among these $N(A)$ repetitions the number of times the event B also occurs (along with A) is $N(AB)$. The ratio $N(AB)/N(A)$ will be called the *conditional frequency ratio of B on the hypothesis that A has occurred* and denoted by $f(B|A)$, i.e.,

$$f(B|A) = \frac{N(AB)}{N(A)} \quad (2.3.8)$$

We assume that $\lim_{N \rightarrow \infty} f(B|A)$ exists, and this limit is called the *conditional probability of B on the hypothesis that A has occurred*, to be denoted by $P(B|A)$. That is,

$$P(B|A) = \lim_{N \rightarrow \infty} f(B|A) \quad (2.3.9)$$

Now

$$f(B|A) = \frac{N(AB)}{N} / \frac{N(A)}{N} = \frac{f(AB)}{f(A)}$$

As $N \rightarrow \infty$ we get

$$P(B|A) = \frac{P(AB)}{P(A)}$$

provided $P(A) \neq 0$.

Similarly

$$P(A|B) = \frac{P(AB)}{P(B)}$$

provided $P(B) \neq 0$.

Hence, if $P(A), P(B) \neq 0$, we have the *multiplication rule* :

$$P(AB) = P(A)P(B|A) = P(B)P(A|B) \quad (2.3.10)$$

Criticisms of the new theory

Although there is not much objection against the logical content of this theory, there is some inherent weakness or inelegance in the mathematical formalism. In this definition, we note that the frequency ratio is thoroughly an empirical concept, whereas the limit is postulated in a rigorous analytical sense. This combination of empirical and theoretical concepts is very inelegant and naturally leads to mathematical difficulties. Now this problem is not typical of probability theory only but arises in other branches of mathematics as well, e.g. theories of geometry. In geometry, we face the same difficult situation if we try to define the fundamental entities like a point, a straight line etc. We may attempt to define a point as the limit of a sequence of chalk dots drawn on the blackboard of gradually decreasing dimensions, which will be somewhat similar to the above definition of probability. This is, however, not done in modern theories of geometry, in which point, straight line etc. remain undefined concepts, and we start with a system of axioms which specify the fundamental relations among them. Mathematicians, as we know, are habitually reluctant to accept things as new and, as such, would always try to define apparently new things in terms of things already known. But suppose if a concept is radically new and can in no way be explained in terms of old ideas, then the

question of a formal definition becomes meaningless. In the theory of probability also, we are ultimately forced to give up the hope of defining probability and take recourse to an axiomatic theory in which probability is accepted as an undefined new concept, and only the salient rules for calculation of probabilities are postulated. These rules will, however, be chosen from the previous theories with necessary modifications for operational convenience. And for this, we go over to the next chapter.



S.C.E.R.T., West Bengal

Date 10-3-87

Acc. No. 3851

CHAPTER 3

FUNDAMENTAL AXIOMS

3.1 AXIOMS OF MATHEMATICAL PROBABILITY

Let E be a random experiment described by the event space S and A be any event connected with E , i.e. $A \subseteq S$. The probability of A is a number associated with A , to be denoted by $P(A; E)$ or simply $P(A)$, such that the following axioms are satisfied :

I. $P(A) \geq 0$

II. The probability of a certain event, $P(S) = 1$.

III. If A_1, A_2, A_3, \dots be a finite or infinite sequence of pairwise mutually exclusive evnts, i.e. $A_i A_j = O$ ($i \neq j$; $i, j = 1, 2, 3, \dots$), then

$$P(A_1 + A_2 + A_3 + \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

This axiom is obviously the formula (2.3.7) of the previous theory with an important extension to an infinite sequence of events and is called the axiom of *complete additivity*.

Frequency interpretation. Now starting from the above axioms, we can logically build up the mathematical structure of the theory of probability. But in order that such a theory may also be meaningful from the point of view of practical applications, we must have to postulate the basic rule for connecting the ideal numbers *probabilities* with experience. This rule, not included in the axioms, consists in the following *frequency interpretation* (not frequency definition!) of probability.

It the random experiment E is repeated a large number of times under identical or uniform conditions, the frequency ratio of any event will be approximately equal to its probability, i.e. $P(A) \simeq f(A)$, so that $f(A)$ can be taken to be an experimentally measured value of the idealised number $P(A)$, and longer is the sequence of repetitions of E more accurate is the measured value.

Remark. In view of the frequency interpretation, we are still

519.9
GUP

restricted to the class of random experiments which can be repeated a large number of times, at least conceptually, under uniform conditions.

Simple deductions

1. From (1.3.1) $A + \bar{A} = S$, $A\bar{A} = O$. Hence

$$1 = P(S) = P(A + \bar{A}) = P(A) + P(\bar{A})$$

or

$$P(\bar{A}) = 1 - P(A) \quad (3.1.1)$$

2. Since $\bar{S} = O$, $P(O) = P(\bar{S}) = 1 - P(S) = 0$, or

$$P(O) = 0 \quad (3.1.2)$$

i.e. the probability of an impossible event is zero.

If, however, $P(A) = 0$, we cannot conclude $A = O$ or A is impossible; in this case, we say that A is *stochastically impossible*. (The word *stochastic* means pertaining to probability.)

Similarly, if $P(A) = 1$, we say that A is *stochastically certain*.

3. $P(A) = 1 - P(\bar{A})$ and since $P(\bar{A}) \geq 0$, we have

$$P(A) \leq 1 \quad (3.1.3)$$

4. Let $A \subseteq B$. Then $B = A + (B - A)$, where A and $B - A$ are mutually exclusive, so that

$$P(B) = P(A) + P(B - A)$$

or

$$P(B - A) = P(B) - P(A) \quad (3.1.4)$$

Further, since $P(B - A) \geq 0$

$$P(A) \leq P(B) \quad (3.1.5)$$

5. CLASSICAL DEFINITION. Let the event space S contain the n points U_1, U_2, \dots, U_n . Then

$$S = U_1 + U_2 + \dots + U_n$$

Since any two event points are necessarily mutually exclusive, $U_i U_j = O$ ($i \neq j$), and so

$$1 = P(S) = P(U_1) + P(U_2) + \dots + P(U_n)$$

or

$$P(U_1) + P(U_2) + \dots + P(U_n) = 1$$

If the event points are assumed to be equally probable, we have

$$P(U_1) = P(U_2) = \dots = P(U_n) \\ = \{P(U_1) + P(U_2) + \dots + P(U_n)\}/n = 1/n$$

If now any event A contains m event points, say, U_1, U_2, \dots, U_m , then

$$A = U_1 + U_2 + \dots + U_m$$

So

$$P(A) = P(U_1) + P(U_2) + \dots + P(U_m) = m/n$$

For clarity, writing $m(A)$ in place of m , we get the classical formula

$$P(A) = \frac{m(A)}{n} \quad (3.1.6)$$

6. GENERAL ADDITION RULE. We shall now extend the addition rule to events which may not be, in general, mutually exclusive. Consider any two events A and B . The events $A - AB$, AB and $B - AB$ are always pairwise mutually exclusive, and we have

$$A = (A - AB) + AB, \quad B = AB + (B - AB)$$

and

$$A + B = (A - AB) + AB + (B - AB)$$

Then

$$P(A) = P(A - AB) + P(AB), \quad P(B) = P(AB) + P(B - AB)$$

$$P(A + B) = P(A - AB) + P(AB) + P(B - AB)$$

Eliminating $P(A - AB)$ and $P(B - AB)$ from the above equations, we get the general addition rule :

$$P(A + B) = P(A) + P(B) - P(AB) \quad (3.1.7)$$

For three events A, B, C

$$\begin{aligned} P(A + B + C) &= P(A) + P(B + C) - P\{A(B + C)\} \\ &= P(A) + P(B) + P(C) - P(BC) - P(AB + AC) \\ &= P(A) + P(B) + P(C) - P(BC) - P(AB) - P(AC) \\ &\quad + P(AB.AC) \end{aligned}$$

Noting that $AB.AC = AABC = ABC$, we have

$$\begin{aligned} P(A + B + C) &= P(A) + P(B) + P(C) - P(BC) - P(CA) \\ &\quad - P(AB) + P(ABC) \end{aligned} \quad (3.1.8)$$

Generalising for n events

$$\begin{aligned}
 P(A_1 + A_2 + \dots + A_n) &= P(A_1) + P(A_2) + \dots + P(A_n) \\
 &\quad - P(A_1 A_2) - P(A_1 A_3) \dots - P(A_{n-1} A_n) \\
 &\quad + P(A_1 A_2 A_3) + P(A_1 A_2 A_4) + \dots + P(A_{n-2} A_{n-1} A_n) + \\
 &\quad \dots + (-1)^{n-1} P(A_1 A_2 \dots A_n)
 \end{aligned} \tag{3.1.9}$$

7. If $\{A_n\}$ is a monotonic sequence of events, then

$$P(\lim A_n) = \lim P(A_n) \tag{3.1.10}$$

Proof. First assume that $\{A_n\}$ is monotonic non-decreasing or expanding, for which

$$\lim A_n = \sum_{n=1}^{\infty} A_n$$

Setting

$$B_1 = A_1, \quad B_n = A_n - A_{n-1} \quad (n \geq 2)$$

$\{B_n\}$ is a sequence of pairwise mutually exclusive events such that

$$\sum_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} B_n$$

Also for every n

$$A_n = \sum_{i=1}^n B_i$$

Using Axiom III we have

$$\begin{aligned}
 P(\lim A_n) &= P\left(\sum_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n) \\
 &= \lim \sum_{i=1}^n P(B_i) = \lim P\left(\sum_{i=1}^n B_i\right) = \lim P(A_n)
 \end{aligned}$$

Next consider the case in which $\{A_n\}$ is a monotonic non-increasing or contracting sequence of events. Then we know that $\{\bar{A}_n\}$ is expanding so that by the above result

$$P(\lim \bar{A}_n) = \lim P(\bar{A}_n)$$

But by (1.3.7)

$$P(\lim \bar{A}_n) = P(\overline{\lim A_n}) = 1 - P(\lim A_n)$$

and $\lim P(\bar{A}_n) = \lim \{1 - P(A_n)\} = 1 - \lim P(A_n)$ whence the result (3.1.10) follows.

Examples

In the following examples we assume that any event space in question is classical in nature, i.e. contains a finite number of event points, all of which are equally probable, unless stated otherwise.

1. A coin is tossed 3 times in succession. Find the probability of (a) 2 heads, (b) 2 consecutive heads.

We have already discussed the event space of this random experiment in Ex. 1, Sec. 1.4. The total number of points in the space, $n=8$. Let A denote the event '2 heads'. Then A contains 3 event points, viz. U_2, U_3, U_4 , i.e. $m(A)=3$. By (3.1.6)

$$P(A) = m(A)/n = 3/8$$

Let B be the event '2 consecutive heads'. B consists of the 2 points U_2 and U_4 so that $m(B)=2$. Hence

$$P(B) = 2/8 = 1/4$$

2. Two dice are thrown. Find the probability that the sum of the faces equals or exceeds 10.

Here $n=36$. Let A, B, C denote the events 'sum 10', 'sum 11' and 'sum 12' respectively. Then $A+B+C$ is the required event, the probability of which is to be computed.

Now the event A contains the 3 points (4, 6), (5, 5), (6, 4), the event B the 2 points (5, 6), (6, 5) and C the only point (6, 6). Then

$$\begin{aligned} m(A) &= 3, & m(B) &= 2, & m(C) &= 1 \\ P(A) &= 3/36, & P(B) &= 2/36, & P(C) &= 1/36 \end{aligned}$$

Since A, B, C are pairwise mutually exclusive, we have by Axiom III

$$P(A+B+C) = P(A) + P(B) + P(C) = 1/6$$

3. A die is rolled. If the result is either an even face or a multiple of three, I win. What is the probability of my winning?

Let A —even face, B —multiple of three.

The event space S and the events A, B are represented by the adjoining diagram (Fig. 2). A and B are not mutually exclusive, but contain one point in common, viz. 6. Clearly then, $n=6$, $m(A)=3$, $m(B)=2$, $m(AB)=1$. So

$$P(A) = 3/6, \quad P(B) = 2/6, \quad P(AB) = 1/6$$

Hence, by (3.1.7), the probability of my win $= P(A+B) = P(A) + P(B) - P(AB) = 2/3$.

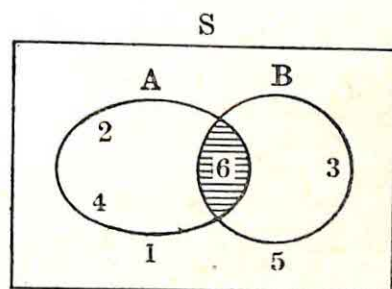


Fig. 2

Another method. We may also directly count the number of points contained in the event 'either an even face or a multiple of three'. It is 4, and hence the required probability is $4/6 = 2/3$.

4. A die is thrown k times in succession. Find the probability of obtaining six at least once.

A typical event point of this space may be represented by a succession of k integers ranging from 1 to 6, say, (5, 3, 1, ..., 4).

The total number of points in the event space is obviously the number of ways in which k places can be filled by 6 different things, repetitions being allowed, and hence $n = 6^k$.

Let A denote the event 'at least one six'. In this case, it will be easier to calculate the probability of the complementary event \bar{A} which is 'no six'. By the same reasoning as above \bar{A} will contain 5^k points, or

$$m(\bar{A}) = 5^k, \quad P(\bar{A}) = (5/6)^k$$

By (3.1.1)

$$P(A) = 1 - (5/6)^k$$

5. A card is drawn at random from each of two well-shuffled packs of cards. What is the probability that at least one of them is queen of spades?

Let A —the first card is queen of spades, and B —the second card is queen of spades. Here $n = 52^2$.

More precisely, A represents the event 'the first card is queen of spades and the second anything', so that $m(A) = 52$. Similarly, $m(B) = 52$.

Therefore

$$P(A) = P(B) = 52/52^2 = 1/52$$

The event AB contains only one point, viz. 'both cards are queens of spades' or $m(AB) = 1$, and so $P(AB) = 1/52^2$.

Now $A + B$ is the required event, and

$$P(A + B) = P(A) + P(B) - P(AB) = \frac{1}{52} + \frac{1}{52} - \frac{1}{52^2} = \frac{103}{2704}$$

Remark. For calculating $P(A)$, we may be tempted to consider the event space of the first draw only which contains a total of 52 points, of which one is contained in A so that $P(A) = 1/52$. The answer is correct, but that would be running intuitively ahead of logic, and the reader is advised to refrain from such intuitive short-cuts at the initial stage, and think clearly in terms of the exact event space of the random experiment in question.

Another method. Let A denote the event 'at least one card is queen of spades'. Then \bar{A} is the event 'none of them is queen of spades', and easily we get $m(\bar{A}) = 51^2$. Hence

$$P(\bar{A}) = \left(\frac{51}{52}\right)^2, \quad P(A) = 1 - \left(\frac{51}{52}\right)^2 = \frac{103}{2704}$$

6. An urn contains $N = N_1 + N_2$ balls, of which N_1 are white and N_2 black. (a) A ball is drawn at random from the urn. What is the probability that it is white? (b) If n balls are drawn, find the probability that among these exactly i balls are white.

(a) The balls of the same colour are assumed to be distinguishable among themselves, and hence when one ball is drawn the total number of points in the event space is N , of which N_1 are contained in the event 'white ball'. Therefore, the probability of drawing a white ball is N_1/N . Similarly, the probability of drawing a black ball is $N_2/N = 1 - N_1/N$.

(b) In this case an event point will be a (disordered) group of n balls, and the total number of event points is the number of different groups of n balls that can be formed out of N different balls, which is $\binom{N}{n}$. If among the n balls drawn i balls are white, the remaining $n - i$ balls must be black. Hence the event ' i white balls' will contain $\binom{N_1}{i} \binom{N_2}{n-i}$ event points, and as such its probability is

$$\frac{\binom{N_1}{i} \binom{N_2}{n-i}}{\binom{N}{n}} \quad (3.1.11)$$

Stirling's formula. We know how laborious it is to calculate the factorials of large numbers, and often it is convenient to have a formula for computing the numerical values of large factorials approximately. This is Stirling's formula which states that

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n+\theta/12n} \quad (0 < \theta < 1) \quad (3.1.12)$$

Since $\theta/12n \rightarrow 0$ as $n \rightarrow \infty$, we have for large values of n

$$n! \simeq \sqrt{2\pi} n^{n+1/2} e^{-n} \quad (3.1.13)$$

This approximation formula is fairly accurate, as we may see that for $n = 10$ the error is 0.8%, and for $n = 100$ it is only 0.08%.

7. What is the probability that a bridge hand of 13 cards contains one ace?

Let us restate the problem in the following manner: A pack contains 52 cards, of which 4 are aces and 48 other cards; 13 cards are drawn from the pack; to find the probability that among these 13 cards one is an ace. Thus we see that this

problem fits in exactly with the model of Ex. 6, and by (3.1.11) the required probability is

$$\binom{4}{1} \binom{48}{12} / \binom{52}{13} \approx 0.44$$

8. The urn problem of Ex. 6 may be easily generalised to balls of more than two different colours. Let the urn contain $N = N_1 + N_2 + \dots + N_m$ balls, of which N_1 are of the first colour, N_2 of the second colour, ..., and N_m of the m th colour. The probability of drawing a ball of the k th colour is then N_k/N ($k = 1, 2, \dots, m$).

If $n = i_1 + i_2 + \dots + i_m$ balls are drawn from the urn, the probability that of these i_1 are of the first colour, i_2 of the second colour, ..., and i_m of the m th colour is, by arguments similar to those in Ex. 6

$$\frac{\binom{N_1}{i_1} \binom{N_2}{i_2} \dots \binom{N_m}{i_m}}{\binom{N}{n}} \quad (3.1.14)$$

9. Find the probability that a bridge hand will contain 5 spades, 4 hearts, 3 diamonds and 1 club.

By (3.1.14) the required probability is

$$\binom{13}{5} \binom{13}{4} \binom{13}{3} \binom{13}{1} / \binom{52}{13} \approx 0.0054$$

10. From an urn containing N_1 white and N_2 black balls ($N = N_1 + N_2$), balls are successively drawn without replacement. What is the probability that i black balls will precede the first white ball?

Suppose all the balls are drawn one by one without replacement and arranged in N different rooms. The total number of event points is the number of ways in which N distinguishable balls can be arranged in N rooms, i.e. $N!$. Now the required event means that the first i rooms are occupied by black balls, the $(i+1)$ th room by a white ball, and the last $N-i-1$ rooms are filled by the remaining $N-i-1$ balls in any manner whatsoever. So the required event contains

$$N_2(N_2-1)\dots(N_2-i+1) \cdot N_1 \cdot (N-i-1)!$$

event points, and hence its probability is

$$\frac{N_1 N_2(N_2-1)\dots(N_2-i+1)}{N(N-1)\dots(N-i)} \quad (3.1.15)$$

Clearly, this result holds for $i \geq 1$. For $i=0$, we may easily get by direct computation that the required probability, which is indeed the probability that the first drawing yields a white ball, is N_1/N .

11. When n dice are thrown, find the probability that the sum of the points on the dice has a prescribed value s .

Clearly the event space contains 6^n event points. The number of event points contained in the required event is the number of different sets of integers (x_1, x_2, \dots, x_n) such that

$$x_1 + x_2 + \dots + x_n = s$$

where x_1, x_2, \dots, x_n can take values $1, 2, \dots, 6$. But this number is again the coefficient of x^s in the expansion of $(x + x^2 + x^3 + x^4 + x^5 + x^6)^n$. We have

$$x + x^2 + x^3 + x^4 + x^5 + x^6 = x(1 - x^6)/(1 - x)$$

and

$$(1 - x^6)^n = \sum_{i=0}^n (-1)^i \binom{n}{i} x^{6i}, \quad (1 - x)^{-n} = \sum_{j=0}^{\infty} \binom{n+j-1}{n-1} x^j$$

Hence

$$(x + x^2 + \dots + x^6)^n = \sum_{i=0}^n \sum_{j=0}^{\infty} (-1)^i \binom{n}{i} \binom{n+j-1}{n-1} x^{n+6i+j}$$

If $n + 6i + j = s$, $j = s - 6i - n$ and as $j \geq 0$, $i \leq (s - n)/6$, and hence the coefficient of x^s in the above expansion is

$$\sum (-1)^i \binom{n}{i} \binom{s-6i-1}{n-1}$$

where i ranges from 0 to the greatest integer $\leq (s - n)/6$. This being also the number of event points the required event contains, its probability is

$$6^{-n} \sum (-1)^i \binom{n}{i} \binom{s-6i-1}{n-1} \quad (3.1.16)$$

12. If r balls are placed at random into n given cells, find the probability that the 1st cell contains r_1 balls, the 2nd cell r_2 balls, and the n th cell r_n balls, where $r_1 + r_2 + \dots + r_n = r$.

The balls, being distinguishable, may be numbered $1, 2, \dots, r$. The total number of points in the event space is n^r , for the 1st ball may be

placed in any one of the n cells, and the same is true for the 2nd, 3rd,..... n th balls. The number of points the required event contains is same as the number of permutations of r distinguishable balls such that the first r_1 balls are placed in the 1st cell, the next r_2 balls in the 2nd cell,..... and the last r_n balls in the n th cell, but, at the same time, we ignore the order of the r_1 balls in the 1st cell, of the r_2 balls in the 2nd cell,..... and of the r_n balls in the n th cell. Hence the number of points contained in the required event is

$$\frac{r!}{r_1! r_2! \dots r_n!}$$

and the required probability is

$$\frac{n^{-r} r!}{r_1! r_2! \dots r_n!} \quad (3.1.17)$$

Indistinguishable balls. In the next two examples we shall consider random distributions of *indistinguishable* balls into a number of cells which serve as important models in studying the behaviour of assemblages in small-particle physics, in which identical particles are represented by the indistinguishable balls and the given cells correspond to the various possible physical states of the particles.

13. If r indistinguishable balls are placed at random into n different cells, find the probability that 1st cell contains r_1 balls, the 2nd cell r_2 balls,..... and the n th cell r_n balls, where $r_1 + r_2 + \dots + r_n = r$.

The event in question is a typical event point which may be represented by the symbol

$$(r_1 \text{ balls} \mid r_2 \text{ balls} \mid \dots \mid r_n \text{ balls})$$

where $r_1 + r_2 + \dots + r_n = r$ and each r_i can take values from 0 to r .

To count how many event points are there in the space, we proceed as follows. In the above symbolical representation of the n cells we have $n-1$ internal partitions and there are r balls, so that the total number of internal partitions and balls is $n+r-1$. Mark the points 1, 2,... $n+r-1$ on the number axis. Any event point may be obtained by choosing $n-1$ of these $n+r-1$ marked points in which the partitions are inserted, the remaining r points being occupied by the r balls. Hence the total number of event points is

$$\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$$

We now give two possible solutions of the above problem.

(a) In the first solution the event points are considered to be *not* equally probable but the probability of an event point is taken to be the probability of the same event if the balls were distinguishable, i.e. the required probability is given by (3.1.17). This is perhaps the most natural choice, and this model in physics is known as the '*Maxwell-Boltzmann statistics*'.

(b) In the second solution all the event points are taken to be equally probable so that the probability of the required event, which is an event point, is

$$\binom{n+r-1}{r}^{-1} \quad (3.1.18)$$

In physics this result goes by the name of '*Bose-Einstein statistics*'.

Remark. It is interesting to note that in the above two different solutions of the problem, different sets of probabilities have been assigned to the same event space giving rise to two different models, both of which are useful in practice.

14. Let r indistinguishable balls be distributed at random to n ($\geq r$) cells such that a cell is either empty or occupied by a single ball (i.e. a cell cannot contain two or more balls). Find the probability that the 1st cell contains r_1 balls, the 2nd cell r_2 balls, and the n th cell r_n balls, where $r_1 + r_2 + \dots + r_n = r$ and each $r_i = 0$ or 1.

The total number of event points is $\binom{n}{r}$, for this is obviously the number of ways of selecting r cells out of the n given cells in which the r balls are placed, the rest of the cells being empty. Now assuming that the event points are equally probable and noting that the required event is a typical event point, its probability is

$$\binom{n}{r}^{-1} \quad (3.1.19)$$

This assignment of probabilities is known as the '*Fermi-Dirac statistics*' in physics.

3.2 CONDITIONAL PROBABILITY

The conditional probability of an event B on the hypothesis that

placed in any one of the n cells, and the same is true for the 2nd, 3rd,..... n th balls. The number of points the required event contains is same as the number of permutations of r distinguishable balls such that the first r_1 balls are placed in the 1st cell, the next r_2 balls in the 2nd cell,..... and the last r_n balls in the n th cell, but, at the same time, we ignore the order of the r_1 balls in the 1st cell, of the r_2 balls in the 2nd cell,..... and of the r_n balls in the n th cell. Hence the number of points contained in the required event is

$$\frac{r!}{r_1! r_2! \dots r_n!}$$

and the required probability is

$$\frac{n^{-r} r!}{r_1! r_2! \dots r_n!} \quad (3.1.17)$$

Indistinguishable balls. In the next two examples we shall consider random distributions of *indistinguishable* balls into a number of cells which serve as important models in studying the behaviour of assemblages in small-particle physics, in which identical particles are represented by the indistinguishable balls and the given cells correspond to the various possible physical states of the particles.

13. If r indistinguishable balls are placed at random into n different cells, find the probability that 1st cell contains r_1 balls, the 2nd cell r_2 balls,..... and the n th cell r_n balls, where $r_1 + r_2 + \dots + r_n = r$.

The event in question is a typical event point which may be represented by the symbol

$$(r_1 \text{ balls} \mid r_2 \text{ balls} \mid \dots \mid r_n \text{ balls})$$

where $r_1 + r_2 + \dots + r_n = r$ and each r_i can take values from 0 to r .

To count how many event points are there in the space, we proceed as follows. In the above symbolical representation of the n cells we have $n-1$ internal partitions and there are r balls, so that the total number of internal partitions and balls is $n+r-1$. Mark the points $1, 2, \dots, n+r-1$ on the number axis. Any event point may be obtained by choosing $n-1$ of these $n+r-1$ marked points in which the partitions are inserted, the remaining r points being occupied by the r balls. Hence the total number of event points is

$$\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$$

We now give two possible solutions of the above problem.

(a) In the first solution the event points are considered to be *not* equally probable but the probability of an event point is taken to be the probability of the same event if the balls were distinguishable, i.e. the required probability is given by (3.1.17). This is perhaps the most natural choice, and this model in physics is known as the '*Maxwell-Boltzmann statistics*'.

(b) In the second solution all the event points are taken to be equally probable so that the probability of the required event, which is an event point, is

$$\left(\frac{n+r-1}{r} \right)^{-1} \quad (3.1.18)$$

In physics this result goes by the name of '*Bose-Einstein statistics*'.

Remark. It is interesting to note that in the above two different solutions of the problem, different sets of probabilities have been assigned to the same event space giving rise to two different models, both of which are useful in practice.

14. Let r indistinguishable balls be distributed at random to $n (\geq r)$ cells such that a cell is either empty or occupied by a single ball (i.e. a cell cannot contain two or more balls). Find the probability that the 1st cell contains r_1 balls, the 2nd cell r_2 balls,..... and the n th cell r_n balls, where $r_1 + r_2 + \dots + r_n = r$ and each $r_i = 0$ or 1.

The total number of event points is $\binom{n}{r}$, for this is obviously the number of ways of selecting r cells out of the n given cells in which the r balls are placed, the rest of the cells being empty. Now assuming that the event points are equally probable and noting that the required event is a typical event point, its probability is

$$\left(\frac{n}{r} \right)^{-1} \quad (3.1.19)$$

This assignment of probabilities is known as the '*Fermi-Dirac statistics*' in physics.

3.2 CONDITIONAL PROBABILITY

The conditional probability of an event B on the hypothesis that

another event A has occurred will be denoted by $P(B|A)$ and defined by

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (3.2.1)$$

provided $P(A) \neq 0$.

In case $P(A) = 0$, the conditional probability $P(B|A)$ remains undefined.

Interpretation. The interpretation of these newly defined numbers—*conditional probabilities* will naturally be as follows. For a long sequence of repetitions of the random experiment E under uniform conditions, the conditional frequency ratio $f(B|A)$ is taken to be an approximate value of the conditional probability $P(B|A)$.

Similarly, by definition

$$P(A|B) = \frac{P(AB)}{P(B)}$$

provided $P(B) \neq 0$. Hence, if $P(A), P(B) \neq 0$, we have

$$P(AB) = P(A)P(B|A) = P(B)P(A|B) \quad (3.2.2)$$

In case the conditional probabilities can be directly computed from the conditions of the experiment, (3.2.2) gives us a formula for calculating the probability of the product of two events and hence is often called the *multiplication rule*.

Remark. It may seem somewhat paradoxical when we say that the same equation (3.2.1) or (3.2.2) is used to define the conditional probabilities and as a multiplication rule. The conditional probabilities are certainly new things in our theory and need be defined, and the latter statement of a multiplication rule is made only in a practical sense. What we mean is that, if the conditional probabilities can be determined in a practical manner by methods other than using the definition itself, then the probability of the joint occurrence of two events may be calculated by formula (3.2.2). This situation is indeed slightly unhappy but has big parallels in other branches of mathematics as well. In mechanics, we remember, the law—*force = mass \times acceleration* is primarily used to measure force but also serves as an equation of motion if the force can be measured by indirect means.

Generalisation. For three events A, B, C , we shall have

$$P(ABC) = P(A) P(B|A) P(C|AB) \quad (3.2.3)$$

Proof. R. H. S. $= P(A) \cdot \frac{P(AB)}{P(A)} \cdot \frac{P(ABC)}{P(AB)} = P(ABC)$

In general, for n events the multiplication rule is

$$P(A_1 A_2 \dots A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1 A_2) \dots P(A_n | A_1 A_2 \dots A_{n-1}) \quad (3.2.4)$$

$$1. \quad P(AB|A) = P(B|A)$$

$$2. \quad P(B + C | A) = P(B | A) + P(C | A) - P(BC | A)$$

If $ABC = O$,

$$P(B + C | A) = P(B | A) + P(C | A)$$

3. If the event space S contains n equally probable event points, then $P(A) = m(A)/n$, $P(AB) = m(AB)/n$. Hence by (3.2.1)

$$P(B|A) = \frac{m(AB)}{m(A)} \quad (3.2.5)$$

Thus when we make the hypothesis that the event A has occurred, we are, so to say, restricted to a new event space A . The portion of B which is left in the new space is evidently AB which contains $m(AB)$ points, while the total number of points in this space is $m(A)$, so that formula (3.2.5) only expresses the classical rule in a modified form.

Examples

1. In Ex. 3, Sec. 3.1 compute the conditional probabilities $P(B|A)$ and $P(A|B)$, where A and B denote the events 'even face' and 'multiple of three' respectively.

We have already found

$$n = 6, m(A) = 3, m(B) = 2, m(AB) = 1$$

Hence

$$P(A) = 3/6 = 1/2, P(B) = 2/6 = 1/3, P(AB) = 1/6$$

By definition

$$P(B|A) = P(AB)/P(A) = 1/3$$

$$P(A|B) = P(AB)/P(B) = 1/2$$

We may also calculate more easily by (3.2.5).

$$P(B|A) = m(AB)/m(A) = 1/3, P(A|B) = m(AB)/m(B) = 1/2$$

2. Two cards are drawn successively from a pack without replacing the first. If the first card is a spade, find the probability that the second card is also a spade.

If A —first card is a spade, B —second card is a spade, then AB —both cards are spades. Thus

$$m(A) = 13 \times 51, \quad m(AB) = 13 \times 12$$

and

$$P(B|A) = \frac{m(AB)}{m(A)} = \frac{13 \times 12}{13 \times 51} = \frac{4}{17}$$

We may also arrive at the same result by the following practical mode of reasoning. When the first card is seen to be a spade, there remain in the pack 51 cards, of which 12 are spades. Hence the probability that the second card is also a spade is $12/51 = 4/17$.

If now we feel that we have learnt how to calculate the conditional probabilities directly, we may attempt to solve the next problem by using the multiplication rule.

3. In Ex. 2 find the probability that both cards are spades.

As before

$$P(A) = 13/52 = 1/4, \quad P(B|A) = 4/17$$

Hence

$$P(AB) = P(A)P(B|A) = 1/17$$

The result may be verified by direct computation by formula (3.1.6).

4. POLYA'S URN PROBLEM. From an urn containing r red and b black balls, n balls are successively drawn such that the ball drawn is always replaced and, in addition, c balls of the colour drawn are added to the urn. Find the probability of a complete run of n black balls.

Let A_i denote the event 'ith ball is black' ($i = 1, 2, \dots, n$). Then the required event is $A_1 A_2 \dots A_n$. Clearly

$$P(A_1) = \frac{b}{r+b}$$

Now make the hypothesis that the event A_1 has occurred, i.e. the 1st ball is black, so that the urn now contains r red and $b+c$ black balls. Hence the conditional probability

$$P(A_2|A_1) = \frac{b+c}{r+b+c}$$

Similarly, it follows that

$$P(A_3 | A_1 A_2) = \frac{b+2c}{r+b+2c}$$

$$\dots \dots \dots$$

$$P(A_n | A_1 A_2 \dots A_{n-1}) = \frac{b+(n-1)c}{r+b+(n-1)c}$$

By (3.2.4)

$$P(A_1 A_2 \dots A_n) = \frac{b(b+c)(b+2c)\dots[b+(n-1)c]}{(r+b)(r+b+c)(r+b+2c)\dots[r+b+(n-1)c]}$$

5. MATCH OR RENCONTRE PROBLEM. From an urn containing n tickets numbered $1, 2, \dots, n$, tickets are drawn successively without replacement. If the k th ticket appears at the k th drawing, then we have a *match* or *rencontre*. Find the probabilities of (a) at least one match, (b) no match at all and (c) exactly i matches.

Let the tickets drawn be arranged in n different rooms. The total number of event points is the number of ways in which n tickets can be arranged in n rooms, which is $n!$.

Let the event A_k be 'match at the k th drawing' ($k=1, 2, \dots, n$). Let us first calculate $P(A_1 A_2 \dots A_k)$. The number of event points which $A_1 A_2 \dots A_k$ contains is $(n-k)!$, since the rooms no. $1, 2, \dots, k$ are filled by tickets no. $1, 2, \dots, k$ respectively so that the remaining $n-k$ rooms can be filled by the remaining $n-k$ tickets in $(n-k)!$ ways. Hence

$$P(A_1 A_2 \dots A_k) = (n-k)! / n!$$

(a) By (3.1.9) and using symmetry, the probability of at least one match

$$P(A_1 + A_2 + \dots + A_n)$$

$$= n P(A_1) - \binom{n}{2} P(A_1 A_2) + \dots + (-1)^{n-1} P(A_1 A_2 \dots A_n)$$

$$= \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} P(A_1 A_2 \dots A_k)$$

$$= \sum_{k=1}^n (-1)^{k-1} / k!$$

(b) Probability of no match at all is

$$1 - P(A_1 + A_2 + \dots + A_n) = \sum_{k=0}^n (-1)^k / k !$$

(c) From symmetry, the probability of exactly i matches = $\binom{n}{i} P(A_1 A_2 \dots A_i B)$, where B is the event—'no match in the last $n-i$ drawings'. Now

$$\begin{aligned} P(A_1 A_2 \dots A_i B) &= P(A_1 A_2 \dots A_i) P(B | A_1 A_2 \dots A_i) \\ &= \frac{(n-i)!}{n!} P(B | A_1 A_2 \dots A_i) \end{aligned}$$

For calculating the conditional probability $P(B | A_1 A_2 \dots A_i)$, we note that if the event $A_1 A_2 \dots A_i$ has occurred, then $n-i$ tickets are left in the urn, viz. tickets no. $i+1, i+2, \dots, n$ and drawings no. $i+1, i+2, \dots, n$ are yet to be made. Hence by replacing n by $n-i$ in the result of case (b).

$$P(B | A_1 A_2 \dots A_i) = \sum_{k=0}^{n-i} (-1)^k / k !$$

Therefore the probability of exactly i matches is

$$\frac{1}{i!} \sum_{k=0}^{n-i} (-1)^k / k ! \quad (3.2.6)$$

Theorem. If A_1, A_2, \dots, A_n be a given set of n pairwise mutually exclusive events, one of which certainly occurs, i.e. $A_i A_j = O$ ($i \neq j$; $i, j = 1, 2, \dots, n$) and $A_1 + A_2 + \dots + A_n = S$, then for any arbitrary event X

$$P(X) = P(A_1)P(X|A_1) + P(A_2)P(X|A_2) + \dots + P(A_n)P(X|A_n) \quad (3.2.7)$$

and (Bayes' theorem) if $P(X) \neq 0$

$$P(A_i | X) = \frac{P(A_i) P(X|A_i)}{P(A_1)P(X|A_1) + P(A_2)P(X|A_2) + \dots + P(A_n)P(X|A_n)} \quad (i = 1, 2, \dots, n) \quad (3.2.8)$$

Proof. We have

$$X = SX = (A_1 + A_2 + \dots + A_n)X = A_1 X + A_2 X + \dots + A_n X$$

Since $(A_i X)(A_j X) = A_i A_j X = OX = O$, ($i \neq j$; $i, j = 1, 2, \dots, n$), $A_1 X, A_2 X, \dots, A_n X$ are pairwise mutually exclusive events, and hence

$$P(X) = P(A_1 X) + P(A_2 X) + \dots + P(A_n X)$$

Since $P(A_i X) = P(A_i) P(X|A_i)$, the result (3.2.7) follows. Again

$$P(A_i X) = P(X) P(A_i|X)$$

Hence if $P(X) \neq 0$

$$P(A_i|X) = \frac{P(A_i) P(X|A_i)}{P(X)}$$

Then using (3.2.7), (3.2.8) is proved.

Remarks.

1. Formula (3.2.7) is useful only when the conditional probabilities $P(X|A_1), P(X|A_2), \dots, P(X|A_n)$ can be more easily obtained than a direct computation of $P(X)$.

2. Formula (3.2.8) is known as Bayes' theorem. If we fancy to call the events A_1, A_2, \dots, A_n *causes* of any event, then one of these causes necessarily acts, and if any event X occurs, it must be due to one of these causes, which is the reading of the equation $X = A_1 X + A_2 X + \dots + A_n X$. Now if the probabilities of the causes $P(A_1), P(A_2), \dots, P(A_n)$ are known, and the probabilities of the occurrence of X on the hypotheses that the different causes are acting, viz. $P(X|A_1), P(X|A_2), \dots, P(X|A_n)$ can be calculated, then on the knowledge that the event X has occurred, Bayes' theorem provides the rule for calculating the probability that a particular cause A_i were acting, i.e. $P(A_i|X)$. Phrased this way, Bayes' theorem appears to be deceptively meaningful, and in old days mathematicians tried to discover many philosophical secrets with the help of this theorem (certainly not making very correct use of it!). In modern thinking, however, all this is nonsense, and Bayes' theorem has no deeper meaning than what is mathematically stated above.

3. The above formulæ (3.2.7) and (3.2.8) may be formally extended, without difficulty, to an infinite sequence of *causes*, $\{A_n\}$.

Examples

6. There are three identical urns containing white and black balls. The first urn contains 2 white and 3 black balls, the second urn 3 white and 5 black balls,

and the third urn 5 white and 2 black balls. An urn is chosen at random, and a ball is drawn from it. If the ball drawn is white, what is the probability that the second urn is chosen?

Let A_i denote the event 'the ball is from the i th urn' ($i=1, 2, 3$). The events A_1, A_2, A_3 are pairwise mutually exclusive, and one of these necessarily occurs. We note, the event A_i may be alternatively titled as 'the i th urn is chosen' when, from symmetry, we can write immediately $P(A_1)=P(A_2)=P(A_3)=1/3$.

Let X denote the event 'white ball'. Then easily we find

$$P(X|A_1)=2/5, P(X|A_2)=3/8, P(X|A_3)=5/7$$

By (3.2.7)

$$P(X)=P(A_1)P(X|A_1)+P(A_2)P(X|A_2)+P(A_3)P(X|A_3)$$

$$= \frac{1}{3} \cdot \frac{2}{5} + \frac{1}{3} \cdot \frac{3}{8} + \frac{1}{3} \cdot \frac{5}{7} = \frac{115}{840}$$

Then by (3.2.8)

$$P(A_2|X) = \frac{P(A_2)P(X|A_2)}{P(X)} = \frac{\frac{1}{3} \cdot \frac{3}{8}}{\frac{115}{840}} = \frac{1 \cdot 3 \cdot 0}{115 \cdot 0}$$

7. LAPLACE'S URN PROBLEM. There are $(N+1)$ identical urns marked $0, 1, 2, \dots, N$, each of which contains N white and black balls. The i th urn contains i black and $N-i$ white balls ($i=0, 1, \dots, N$). An urn is chosen at random, and n random drawings are made from it, the ball drawn being always replaced. If all the n balls turn out to be black, what is the probability that the next ball drawn will also be black?

Let A_i denote the event 'the i th urn is chosen'. Then the events A_0, A_1, \dots, A_N are pairwise mutually exclusive, one of which certainly occurs. From symmetry, $P(A_i)=1/(N+1)$. Let X be the event 'all n balls are black'.

By reasoning similar to that in Ex. 4, $P(X|A_i)=i^n/N^n$, and by (3.2.7)

$$P(X) = \sum_{i=0}^N P(A_i) P(X|A_i) = \frac{1}{N+1} \sum_{i=0}^N \left(\frac{i}{N}\right)^n$$

If Y - the $(n+1)$ th ball is black, then XY - all the $n+1$ balls are black. Replacing n by $n+1$ in the above result

$$P(XY) = \frac{1}{N+1} \sum_{i=0}^N \left(\frac{i}{N}\right)^{n+1}$$

Hence the required conditional probability is

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{\sum_{i=0}^N \left(\frac{i}{N}\right)^{n+1}}{\sum_{i=0}^N \left(\frac{i}{N}\right)^n}$$

The result assumes a simple interesting form if N is very large. In that case

$$P(X) \simeq \frac{1}{N} \sum_{i=0}^N \left(\frac{i}{N}\right)^n \simeq \int_0^1 x^n dx = \frac{1}{n+1}$$

$$P(XY) \simeq \frac{1}{n+2}, \quad P(Y|X) \simeq \frac{n+1}{n+2}$$

This is called the *law of succession* of Laplace.

3.3 STOCHASTIC INDEPENDENCE

If $P(B|A) = P(B)$, we may say that the information that the event A has occurred does not affect the probability of the event B , and B is then said to be stochastically independent of A . It follows from (3.2.2) that if $P(B|A) = P(B)$, then $P(A|B) = P(A)$, i.e. A is stochastically independent of B and $P(AB) = P(A)P(B)$, provided, of course, $P(A), P(B) \neq 0$. Also the last equation implies both $P(B|A) = P(B)$ and $P(A|B) = P(A)$.

Thus we define two events A and B to be *stochastically independent* or simply *independent* if

$$P(AB) = P(A)P(B) \tag{3.3.1}$$

We agree to accept this definition even if $P(A)$ or $P(B) = 0$.

Formula (3.3.1) may be used as a simple multiplication rule if we can judge a priori that the events A and B are independent. In a practical problem, two events may be taken to be stochastically independent if there is no *causal* relation between them, i.e. if they are *causally independent*.

Generalisation. Take three events A, B, C which are pairwise independent, i.e.

$$\begin{aligned} P(AB) &= P(A)P(B), \quad P(BC) = P(B)P(C) \\ P(CA) &= P(C)P(A) \end{aligned} \quad (3.3.2)$$

Then we may be led to think that perhaps in this case it follows that the events A and BC etc. are also independent. That such a conjecture is false will be apparent from the following example.

Example 1. Let a coin be tossed twice so that the event space consists of the four points $(H, H), (H, T), (T, H)$ and (T, T) .

Define events A, B, C to be 'head in the first toss', 'head in the second toss' and 'one head' respectively. Then A contains the two points $(H, H), (H, T)$, B the two points $(H, H), (T, H)$ and C the two points $(H, T), (T, H)$. AB contains the only point (H, H) , BC the point (T, H) and CA the point (H, T) , and the event ABC is impossible. Hence

$$\begin{aligned} P(A) &= P(B) = P(C) = 1/2 \\ P(AB) &= P(BC) = P(CA) = 1/4 \\ P(ABC) &= 0 \end{aligned}$$

so that conditions (3.3.2) are satisfied but $P(ABC) \neq P(A)P(BC)$, i.e. A and BC are not independent.

Now suppose that A, B, C are pairwise independent events, and further that the events A and BC etc. are independent. Then (3.3.2) is satisfied and

$$P(ABC) = P(A)P(BC)$$

or by (3.3.2)

$$P(ABC) = P(A)P(B)P(C) \quad (3.3.3)$$

Conversely, if the conditions (3.3.2) and (3.3.3) are both satisfied, then A, B, C are obviously pairwise independent, and

$$P(ABC) = P(A)P(B)P(C) = P(A)P(BC)$$

so that A and BC are independent. Similarly, it follows that B and CA are independent, as also C and AB .

These considerations lead to the following enlargement of the concept of independence for three events.

Remark. Since the die is thrown independent of the toss, it is natural to expect that the events 'head' and 'six' are independent. The independence of the toss and the throw of the die is, however, only meant in the intuitive sense, for we have not yet defined such independence of two random experiments. The only thing we have assumed in the above proof is that the 12 event points are equally probable.

3.4 EXERCISES

1. Prove the following relations :

$$P(\overline{A+B}) = 1 - P(\overline{AB}), \quad P(\overline{AB}) = 1 - P(A) - P(B) + P(AB)$$

$$P(\overline{A+B}) = 1 - P(A) + P(AB), \quad P(\overline{AB}) = P(B) - P(AB)$$

2. Show that the probability of occurrence of only one of the events A and B is $P(A) + P(B) - 2P(AB)$.

3. *Boole's Inequality.* Prove that

$$P(A_1 + A_2 + \dots + A_n) \leq P(A_1) + P(A_2) + \dots + P(A_n)$$

4. A coin is tossed n times in succession. Find the probability of r ($< n$) heads.

5. What is the probability of an odd sum when two dice are thrown ?

6. Two cards are drawn from a well-shuffled pack. Find the probability that at least one of them is a spade.

7. Two urns contain respectively 3 white, 7 red, 15 black balls, and 10 white, 6 red, 9 black balls. One ball is drawn from each urn. Find the probability that both the balls are of the same colour.

8. An urn contains three balls numbered 1, 2 and 3. Two balls are drawn successively, the first ball drawn being replaced. Find the probability that the sum of the two numbers is 5.

9. *De Mere's Paradox.* Show that the probability of obtaining six at least once in 4 throws with a die is slightly greater than $\frac{1}{2}$, and that of obtaining double six at the least once in 24 throws with two dice is slightly less than $\frac{1}{2}$.

10. Find the minimum number of times a die has to be thrown such that the probability of no six is less than $\frac{1}{2}$.

11. The numbers 1, 2, ..., n are arranged in random order. What is the probability that the numbers 1 and 2 are always together ?

12. From the numbers 1, 2, ..., $2n+1$ three are chosen at random. Prove that the probability that these are in arithmetical progression is $3n/(4n^2 - 1)$.

13. A coin is tossed $m+n$ times ($m > n$). Show that the probability of exactly m consecutive heads is $(n+3)/2^{m+2}$, and that of at least m consecutive heads is $(n+2)/2^{m+1}$.

14. From an urn containing n balls any number of balls are drawn. Show that the probability of drawing an even number of balls is $(2^{n-1} - 1)/(2^n - 1)$.

15. If an even number of cards are drawn from a full pack, find the probability that these consist half of red and half of black.
16. What is the probability that a bridge hand will contain at least one ace?
17. What is the probability that the combined bridge hands of 'north' and 'south' contain all the 4 aces?
18. 100 prizes will be given in a lottery of 10,000 tickets. Find the minimum number of tickets a person has to buy in order that the probability of his winning at least one prize is greater than $\frac{1}{2}$.
19. Find the probability that a bridge hand contains all 13 face values.
20. If four persons are selected at random from a group of 3 men, 2 women and 4 children, what is the probability that among these there are 1 man, 1 woman and 2 children?
21. What is the probability that a bridge hand contains 5 cards of some suit, 4 of another, 3 of a third, and 1 of the last suit?
22. From an urn containing n tickets numbered 1, 2, ..., n , r tickets are drawn simultaneously and arranged in increasing order of their numbers: $x_1 < x_2 < \dots < x_r$. Show that the probability that $x_i = s$ is

$$\frac{\binom{s-1}{i-1} \binom{n-s}{r-i}}{\binom{n}{r}}$$

23. An urn contains N_1 white and N_2 black balls. Two players A and B alternately draw a ball without replacement, and one who draws the first white ball wins the game. If A begins to draw, find the probability of his winning.
24. If cards are successively drawn without replacement from a full pack, what is the probability that five cards will precede the first ace?
25. An urn contains N_1 white and N_2 black balls, from which k balls are drawn one by one without replacement and laid aside, their colour being unnoted. Then one more ball is drawn. Find the probability that it is white.
26. From an urn containing N_1 white and N_2 black balls ($N = N_1 + N_2$), balls are successively drawn without replacement until only those of the same colour are left. Prove that the probability that the balls left are white is N_1/N .
27. Find the probability of obtaining 14 with 3 dice, and show that it is the same with 5 dice.
28. If r balls are distributed at random in n cells, prove that the probability p_i that a given cell contains exactly i balls is given by

$$p_i = \binom{r}{i} (n-1)^{r-i} / n^r$$

Further show that the most probable number(s) i_m of balls in a given cell is determined by the inequalities

$$(r+1)/n - 1 \leq i_m \leq (r+1)/n$$

i.e. if $(r+1)/n$ is not an integer, i_m = the greatest integer less than $(r+1)/n$, and if $(r+1)/n$ is an integer, $i_m = (r+1)/n - 1$ or $(r+1)/n$.

29. If n objects are distributed at random among a men and b ($< a$) women, then show that the probability that the women get an odd number of objects is $\frac{1}{2}[(a+b)^n - (a-b)^n] / (a+b)^n$.

30. Let r indistinguishable particles be placed at random into n cells. If the particles obey 'Bose-Einstein statistics', prove that the probability that there are exactly i particles in a given cell is

$$\binom{n+r-i-2}{r-i} / \binom{n+r-1}{r}$$

Show also that the most probable number of particles in a given cell is zero, provided $n > 2$.

31. If r indistinguishable particles obeying 'Fermi-Dirac statistics' is placed at random into n cells, prove that the probability that a given cell is empty is $1 - r/n$.

32. An urn contains 4 white and 6 black balls. Two balls are successively drawn from the urn without replacement of the first ball. If the first ball is seen to be white, what is the probability that the second ball is also white?

33. A secretary writes four letters and the corresponding addresses on envelopes. If he inserts the letters in the envelopes at random irrespective of the addresses, find the probability that only one letter is placed in the corresponding envelope. Also calculate the probability that all the letters are wrongly placed.

34. Ten students have identical raincoats which they hang on the same rack while attending class. After the class each student selects a raincoat at random and goes home. What is the probability that at least one raincoat goes to its original owner?

35. Two urns contain respectively 2 white and 1 black balls, and 1 white and 5 black balls. One ball is transferred from the first to the second urn, and then a ball is drawn from the second urn. What is the probability that the ball drawn is white?

36. There are two identical urns containing respectively 4 white and 3 red balls and 3 white and 7 red balls. An urn is chosen at random, and a ball is drawn from it. Find the probability that the ball is white. If the ball drawn is white, what is probability that it is from the first urn?

37. There are three identical boxes, each provided with two drawers. In the first, each drawer contains a gold coin; in the third, each drawer contains a silver coin; and in the second, one drawer contains a gold and the other a silver coin. A box is selected at random, and one of the drawers is opened. If a gold coin is found, what is the probability that the box chosen is the second one?

38. Three urns contain respectively 1 white and 2 black balls; 2 white and 1 black balls; 2 white and 2 black balls. One ball is transferred from the first to the second urn; then one ball is transferred from the second to the third urn; finally one ball is drawn from the third urn. Find the probability that the ball is white.

39. There are n urns each containing N balls, of which N_1 are white and N_2 black. One ball is transferred from the 1st to the 2nd urn; then one ball is transferred from the 2nd to the 3rd urn and so on; finally one ball is drawn from the n th urn. Prove that the probability that the ball is white is N_1/N .

40. If two events A and B are independent, show that A and \bar{B} are independent, and hence that \bar{A} and \bar{B} are also independent.

41. Let A, B, C be mutually independent events. Then prove that A and $B+C$ are independent and also that \bar{A}, \bar{B}, C are mutually independent.

42. If the probabilities of n mutually independent events be p_1, p_2, \dots, p_n , then show that the probability that at least one of the events will occur is $1 - (1-p_1)(1-p_2) \dots (1-p_n)$.

43. The outcome of an experiment is equally likely to be one of the four points in three-dimensional space with rectangular co-ordinates $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ and $(1, 1, 1)$. If A, B, C denote the events x -co-ordinate 1, y -co-ordinate 1, z -co-ordinate 1 respectively, then check if A, B, C are mutually independent.

COMPOUND EXPERIMENTS

4.1 CARTESIAN PRODUCT OF SETS

Let S and T be any two sets. The cartesian product of S and T , denoted by $S \times T$, is defined to be the set of all *ordered pairs* (x, y) where $x \in S$ and $y \in T$.

For example, if S and T are finite sets such that S contains the m elements x_1, x_2, \dots, x_m and T contains the n elements y_1, y_2, \dots, y_n , then $S \times T$ contains the mn ordered pairs

$$(x_i, y_j) \quad (i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n)$$

We shall also write

$$S \times S = S^2, \quad S \times S \times S = S^3 \text{ etc.}$$

4.2 JOINT INDEPENDENT EXPERIMENTS

Let E be a random experiment described by the event space S which contains m event points, viz. U_1, U_2, \dots, U_m having given probabilities

$$P(U_i) = p_i \quad (i = 1, 2, \dots, m) \quad (4.2.1)$$

so that

$$\sum_{i=1}^m p_i = 1 \quad (4.2.2)$$

Let E' be another experiment and S' its event space containing the n points U'_1, U'_2, \dots, U'_n with probabilities

$$P(U'_j) = p'_j \quad (j = 1, 2, \dots, n) \quad (4.2.3)$$

such that

$$\sum_{j=1}^n p'_j = 1 \quad (4.2.4)$$

The experiments E and E' are performed successively in such a manner that the second experiment E' is independent of the result of the first experiment E , i.e. the result of E does not in any way affect the performance of E' . We are, in fact, going to define the independence

of the experiments E and E' , but for motivating the same we argue *intuitively* as follows.

The joint performance of E and E' we shall call the compound experiment E'' . The event points connected with E'' will then be

$$(U_i, U_j) \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

so that the corresponding event space $S'' = S \times S'$.

Now consider the two events ' U_i occurs in E ' and ' U_j ' occurs in E' ', both connected with E'' . The former contains the event points

$$(U_i, U_1), (U_i, U_2), \dots, (U_i, U_n)$$

and the latter the points

$$(U_1, U_j), (U_2, U_j), \dots, (U_m, U_j)$$

so that their product is the event point (U_i, U_j) . Since the first experiment is independent of the second experiment, the event ' U_i occurs in E ' connected with E'' may be simply regarded as the event point U_i connected with E , which has a probability p_i . Similarly, the event ' U_j ' occurs in E' ' has a probability p_j . Now since the experiments are independent, it is reasonable to assume the two events ' U_i occurs in E ' and ' U_j ' occurs in E' ' to be stochastically independent, and hence

$$P\{(U_i, U_j)\} = p_i p_j \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n) \quad (4.2.5)$$

We now define the experiments E and E' to be *independent* if the assignment of probabilities to the different event points of S'' is given by (4.2.5). We have

$$S'' = \sum_i \sum_j (U_i, U_j)$$

So

$$\begin{aligned} P(S'') &= \sum_i \sum_j P\{(U_i, U_j)\} = \sum_i \sum_j p_i p_j \\ &= (\sum_i p_i) (\sum_j p_j) = 1.1 = 1 \end{aligned}$$

which is a necessary condition.

Theorem. If A and B are any two events connected with E and E' respectively and E, E' are independent, then

$$P\{(A, B)\} = P(A) P(B) \quad (4.2.6)$$

Proof. Let

$$A = \sum_{\alpha} U_{\alpha}, \quad B = \sum_{\beta} U_{\beta}$$

where the indices α and β run over some subsets of the sets $1, 2, \dots, m$ and $1, 2, \dots, n$ respectively. Then

$$P(A) = \sum_{\alpha} P(U_{\alpha}) = \sum_{\alpha} p_{\alpha}, \quad P(B) = \sum_{\beta} P(U_{\beta}) = \sum_{\beta} p_{\beta}'$$

The event (A, B) connected with E'' may be written as

$$(A, B) = \sum_{\alpha} \sum_{\beta} (U_{\alpha}, U_{\beta})$$

Hence

$$\begin{aligned} P\{(A, B)\} &= \sum_{\alpha} \sum_{\beta} P\{(U_{\alpha}, U_{\beta})\} = \sum_{\alpha} \sum_{\beta} p_{\alpha} p_{\beta}' \\ &= \left(\sum_{\alpha} p_{\alpha}\right) \left(\sum_{\beta} p_{\beta}'\right) = P(A) P(B) \end{aligned}$$

The above formulation may be easily generalised to more than two experiments.

4.3 REPEATED INDEPENDENT TRIALS

If the experiment E is itself repeated twice, then the compound experiment E_2 will have event space $S \times S = S^2$ which contains the m^2 points (U_i, U_j) ($i, j = 1, 2, \dots, m$). The independence of two trials will be realised in practice if they are performed under *identical conditions* and mathematically defined by

$$P\{(U_i, U_j)\} = p_i p_j \quad (i, j = 1, 2, \dots, m) \quad (4.3.1)$$

which follows from (4.2.5).

For r independent trials of E , the compound experiment will be denoted by E_r , the corresponding event space being S^r . There are m^r points in S^r , viz.

$$(U_{i_1}, U_{i_2}, \dots, U_{i_r}) \quad (i_1, i_2, \dots, i_r = 1, 2, \dots, m)$$

their probabilities being given by

$$\begin{aligned} P\{(U_{i_1}, U_{i_2}, \dots, U_{i_r})\} &= p_{i_1} p_{i_2} \dots p_{i_r} \\ &\quad (i_1, i_2, \dots, i_r = 1, 2, \dots, m) \quad (4.3.2) \end{aligned}$$

The generalisation of the last theorem will be

Theorem. Let A_1, A_2, \dots, A_r be any events connected with the random experiment E . Then for r independent trials of E

$$P\{(A_1, A_2, \dots, A_r)\} = P(A_1) P(A_2) \dots P(A_r) \quad (4.3.3)$$

Example. From an urn containing n tickets numbered $1, 2, \dots, n$, k tickets are drawn at a time and replaced before the next drawing. Find the probability that in r such drawings, tickets no. $1, 2, \dots, r$ do not appear in the 1st, 2nd, \dots , r th drawings respectively.

Let A_i be the event that the i th ticket does not appear in a single drawing of k tickets ($i = 1, 2, \dots, r$). Then

$$P(A_i) = \frac{\binom{n-1}{k}}{\binom{n}{k}} = \frac{n-k}{n}$$

Since the balls drawn are always replaced, we have r independent trials of the above experiment, and the required event is (A_1, A_2, \dots, A_r) so that by (4.3.3)

$$P\{(A_1, A_2, \dots, A_r)\} = (n-k)^r / n^r$$

4.4 BERNOULLI TRIALS

If a random experiment be such that its event space consists of only two points which are usually called 'success' and 'failure', then a sequence of independent trials of the experiment will be called a *Bernoullian sequence of trials*, provided the probability of 'success' (or 'failure') remains the same for all trials.

Let E be the given random experiment, its event space S containing the two points 'success' and 'failure', to be denoted by the symbols s and f respectively. Let p be the probability of success, i.e.

$$P(s) = p, \quad P(f) = 1 - p = q \text{ (say)}$$

The compound experiment of n independent trials of E , denoted by E_n , has event space S^n containing 2^n points represented by successions of the symbols s and f of the type (s, s, f, s, \dots, f) . By (4.3.2) their probabilities are given by

$$P\{(s, s, f, s, \dots, f)\} = p.p.q.p \dots q \quad (4.4.1)$$

Binomial law. Let A_i denote the event ' i successes' (consequently $n-i$ failures) connected with the compound experiment E_n . Let the

n trials be represented by n rooms. Then the required event means that of these n rooms i rooms are selected in which we place the symbols s , the remaining $n-i$ rooms being filled by the symbols f . Since i rooms can be chosen out of n rooms in $\binom{n}{i}$ different ways, the event A_i contains $\binom{n}{i}$ event points each having probability $p^i q^{n-i}$, so that

$$P(A_i) = \binom{n}{i} p^i q^{n-i}$$

$$= \binom{n}{i} p^i (1-p)^{n-i} \quad (i = 0, 1, 2, \dots, n) \quad (4.4.2)$$

This is called the *binomial law*.

1. The events A_0, A_1, \dots, A_n are pairwise mutually exclusive, one of which necessarily occurs, i.e.

$$S^n = A_0 + A_1 + \dots + A_n$$

So

$$P(S^n) = \sum_{i=0}^n P(A_i) = \sum_{i=0}^n \binom{n}{i} p^i q^{n-i} = (p+q)^n = 1$$

2. $P(A_0) = q^n$, which is the probability of no success at all. Hence the probability of at least one success $= P(\bar{A}_0) = 1 - q^n$. And $P(A_n) = p^n$, i.e. the probability of a complete run of successes is p^n .

3. Next we prove the following identity :

$$\sum_{i=k}^n \binom{n}{i} p^i q^{n-i} = \frac{\int_0^p x^{k-1} (1-x)^{n-k} dx}{\int_0^1 x^{k-1} (1-x)^{n-k} dx} \quad (4.4.3)$$

where the L.H.S. clearly represents the probability of at least k successes in n trials.

Denoting the R.H.S. by I_k and noting that

$$\int_0^1 x^{k-1} (1-x)^{n-k} dx = B(k, n-k+1) = (k-1)! (n-k)! / n!$$

$$I_k = \frac{n!}{(k-1)!(n-k)!} \int_0^p x^{k-1} (1-x)^{n-k} dx$$

$$= \frac{n!}{k!(n-k)!} \left[p^k (1-p)^{n-k} + (n-k) \int_0^p x^k (1-x)^{n-k-1} dx \right]$$

or

$$I_k - I_{k+1} = \binom{n}{k} p^k q^{n-k}$$

Replacing k by $k+1, k+2, \dots, n-1$ and adding all these results and noting that $I_n = p^n$, the identity (4.4.3) follows.

Remark. We know $B(l, m) = \int_0^1 x^{l-1} (1-x)^{m-1} dx$ is the beta function, and $B_x(l, m) = \int_0^x x^{l-1} (1-x)^{m-1} dx$ is called the *incomplete beta function*. Tables of $B_x(k, n-k+1)/B(k, n-k+1)$ have been prepared for different values of p, n, k , from which the individual terms $\binom{n}{i} p^i q^{n-i}$ of the binomial law may be easily obtained.

4. GREATEST TERM. Let $P(A_i)$ be maximum when $i = i_m$, so that i_m may be called the *most probable number of successes*. We have

$$\frac{P(A_i)}{P(A_{i+1})} - 1 = \frac{i - (n+1)p + 1}{(n-i)p}$$

If $(n+1)p$ is not an integer, let r denote the greatest integer less than $(n+1)p$, i.e.

$$(n+1)p - 1 < r < (n+1)p$$

Then $P(A_0) < P(A_1) < \dots < P(A_r) > P(A_{r+1}) > \dots > P(A_n)$ which shows that $i_m = r$.

If $(n+1)p$ is an integer, writing $s = (n+1)p$, we have

$$P(A_0) < P(A_1) < \dots < P(A_{s-1}) = P(A_s) > P(A_{s+1}) > \dots > P(A_n)$$

so that $i_m = s-1$ or s .

The results of both the cases may be combined into the single statement that i_m is the integer(s) determined by the following inequalities :

$$(n+1)p - 1 \leq i_m \leq (n+1)p \quad (4.4.4)$$

Examples

1. In Ex. 1, Sec. 3.1 we can find the probability of 2 heads directly by the binomial law (4.4.2). Our random experiment now consists in tossing the coin and observing if it is head, so that the event 'head' may be called a success and 'tail' a failure. Here the probability of success, $p = \frac{1}{2}$ so that $q = 1 - \frac{1}{2} = \frac{1}{2}$. Hence the required probability is

$$P(A_2) = \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) = \frac{3}{8}$$

2. Take Ex. 4, Sec. 3.1. If the experiment consists in throwing the die and seeing if the result is six, the probability of success, $p = 1/6$ and $q = 5/6$. Therefore the probability of least one six in k throws with the die $= 1 - (5/6)^k$.

3. A and B play a game which must be either won or lost. If the probability that A wins any game is p , find the probability that A wins m games before B wins n games ($m, n \geq 1$).

The issue will be clearly decided in $m+n-1$ games, and the required event is equivalent to, i.e. implies and is implied by, the event that of these $m+n-1$ games A wins at least m games. Now calling A 's winning a game a success, we have $m+n-1$ Bernoulli trials with probability of success p , and the required event being at least m successes, its probability, using (4.4.3), is

$$\sum_{i=m}^{m+n-1} \binom{n}{i} p^i (1-p)^{n-i} = \frac{\int_0^p x^{m-1} (1-x)^{n-1} dx}{\int_0^1 x^{m-1} (1-x)^{n-1} dx}$$

4. DRAWINGS WITH REPLACEMENT. Let an urn contain $N = N_1 + N_2$ balls, of which N_1 are white and N_2 black. If n balls are drawn successively from the urn, the ball drawn being replaced each time, find probability that i drawings will yield white balls.

We know the probability of drawing a white ball $= N_1/N$. Consider the experiment of drawing a ball from the urn and noting if its colour is white. Then the probability of success, $p = N_1/N$ and $q = 1 - p = N_2/N$. Now the random experiment in question may be regarded as a Bernoullian sequence of n trials of the above experiment, and hence the probability of i white balls is

$$P(A_i) = \binom{n}{i} p^i q^{n-i} = \binom{n}{i} \left(\frac{N_1}{N}\right)^i \left(\frac{N_2}{N}\right)^{n-i} \quad (4.4.5)$$

DRAWINGS WITHOUT REPLACEMENT. If the balls are drawn without replacements, or, in other words, all the n balls are drawn simultaneously, the problem becomes identical with Ex. 6, Sec. 3.1, where we found the probability of i white balls to be

$$\frac{\binom{N_1}{i} \binom{N_2}{n-i}}{\binom{N}{n}}$$

It will be interesting to find the limiting form of the above expression as $N \rightarrow \infty$, keeping p fixed and hence $N_1, N_2 \rightarrow \infty$.

The expression

$$\begin{aligned} &= \frac{N_1(N_1-1)\dots(N_1-i+1)}{i!} \times \frac{N_2(N_2-1)\dots(N_2-n+i+1)}{(n-i)!} \\ &\quad \times \frac{n!}{N(N-1)\dots(N-n+1)} \\ &= \binom{n}{i} \left(\frac{N_1}{N}\right)^i \left(\frac{N_2}{N}\right)^{n-i} \\ &\quad \times \frac{\left(1-\frac{1}{N_1}\right) \dots \left(1-\frac{i-1}{N_1}\right) \left(1-\frac{1}{N_2}\right) \dots \left(1-\frac{n-i-1}{N_2}\right)}{\left(1-\frac{1}{N}\right) \dots \left(1-\frac{n-1}{N}\right)} \\ &\rightarrow \binom{n}{i} p^i q^{n-i} \quad \text{as } N \rightarrow \infty \end{aligned}$$

Thus for drawings without replacement, we get the binomial law in the limiting case.

5. BANACH'S MATCH-BOX PROBLEM. A mathematician always carries two match-boxes, each containing n matches. Whenever he needs, he chooses a box at random and draws a match from it. Find the probability that when the first box is found to be empty for the first time, the second box will contain exactly i matches.

The event in question means that the first box is chosen $n+1$ times corresponding to the n matches drawn from it and the case when it is found empty, and the second box is chosen $n-i$ times so that i matches are left in it, and further in the last drawing the first box is

chosen when it is found empty, but in the remaining drawings the two boxes may be chosen in arbitrary order. Suppose we consider the experiment of choosing a box at random and call the choice of the first box a success. Thus we have a Bernoullian sequence of $n+1+n-i=2n-i+1$ trials with probability of success $p=\frac{1}{2}$, and the probability of occurrence of n successes in the first $2n-i$ trials (the successes appearing in any order) and a success in the last trial is, by (4.3.3) and (4.4.2)

$$\binom{2n-i}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{n-i} \cdot \frac{1}{2} = \binom{2n-i}{n} \left(\frac{1}{2}\right)^{2n-i+1} \quad (4.4.6)$$

which is the required answer.

6. An urn contains n tickets numbered 1 to n , from which a ticket is drawn and replaced r times. What is the probability that the greatest number drawn is i ?

Let X_i denote the event that the greatest number drawn is less than or equal to i , which is identical with the event that each of the r drawings yields a ticket whose number is less than or equal to i . Consider the random experiment of drawing a ticket from the urn and observing if its number is less than or equal to i , i.e. one of the numbers $1, 2, \dots, i$, so that the probability of success is i/n . Since X_i is same as the event of all successes in r Bernoulli trials, $P(X_i) = (i/n)^r$.

Now the required event is obviously $X_i - X_{i-1}$ where $X_{i-1} \subseteq X_i$, so that by (3.1.4) the required probability is

$$\begin{aligned} P(X_i - X_{i-1}) &= P(X_i) - P(X_{i-1}) \\ &= [i^r - (i-1)^r]/n^r \end{aligned} \quad (4.4.7)$$

Poisson approximation

Set $p = \mu/n$, where μ is a given positive number, and we pass to limit as $n \rightarrow \infty$ and hence $p \rightarrow 0$. We have

$$P(A_i) = \binom{n}{i} p^i (1-p)^{n-i} = \frac{n(n-1)\dots(n-i+1)}{i!} p^i (1-p)^{n-i}$$

$$= \frac{\mu^i}{i!} \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{i-1}{n}\right) \frac{\left\{ \left(1 - \frac{\mu}{n}\right)^{-\frac{n}{\mu}} \right\}^{-\mu}}{\left(1 - \frac{\mu}{n}\right)^i}$$

$$\rightarrow e^{-\mu} \frac{\mu^i}{i!} \quad \text{as } n \rightarrow \infty$$

Hence, if the probability of success p is small and the number of trials n large, such that $\mu = np$ is of moderate magnitude, we have the approximation formula

$$P(A_i) \simeq e^{-\mu} \frac{\mu^i}{i!} \quad (i = 0, 1, 2, \dots) \quad (4.4.8)$$

This is called the *Poisson approximation to the binomial law*.

We note, if S^∞ denotes the limiting event space

$$S^\infty = A_0 + A_1 + A_2 + \dots \text{to } \infty$$

Since A_0, A_1, A_2, \dots are pairwise mutually exclusive, by Axiom III

$$P(S^\infty) = P(A_0) + P(A_1) + P(A_2) + \dots$$

$$= e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i}{i!} = e^{-\mu} \cdot e^{\mu} = 1$$

Example 7. An urn contains 1 white and 99 black balls. If 1000 drawings are made with replacements, what is the probability of 10 white balls?

Using the notations of Ex. 4, $p = 1/100$, $n = 1000$, so that $\mu = np = 10$.

Here p is small and n large such that μ is of moderate magnitude, and hence we can conveniently use the Poisson approximation (4.4.8). Therefore the required probability $\simeq e^{-10} 10^{10}/10! = 0.125$.

The result given by the binomial law is 0.126, which shows that Poisson approximation is fairly good.

4.5 POISSON TRIALS

A sequence of independent trials of a random experiment, the event space of which contains two points—success and failure, is called a *Poisson sequence of trials* if the probability of success is not constant but varies from one trial to another. It will be rather cumbersome to deduce general formulæ for a Poisson sequence of n trials, and let us, for convenience, consider only three trials. Let the probabilities of success be p_1, p_2, p_3 in the three trials respectively, and hence the probabilities of failure are respectively $1 - p_1 = q_1, 1 - p_2 = q_2, 1 - p_3 = q_3$. The event space of the compound experiment will then contain the following 8 points:

$$\begin{array}{lll} U_1 = (S, S, S), & U_2 = (S, S, F), & U_3 = (S, F, S), \\ U_4 = (F, S, S), & U_5 = (F, F, S), & U_6 = (F, S, F), \\ & U_7 = (S, F, F), & U_8 = (F, F, F) \end{array}$$

Since the trials are independent, the distribution of probabilities in this event space is given by

$$\begin{aligned} P(U_1) &= p_1 p_2 p_3, & P(U_2) &= p_1 p_2 q_3, & P(U_3) &= p_1 q_2 p_3, \\ P(U_4) &= q_1 p_2 p_3, & P(U_5) &= q_1 q_2 p_3, & P(U_6) &= q_1 p_2 q_3, \\ P(U_7) &= p_1 q_2 q_3, & P(U_8) &= q_1 q_2 q_3 \end{aligned}$$

If A_i denotes the event ' i successes', then

$$A_0 = U_8, \quad A_1 = U_5 + U_6 + U_7, \quad A_2 = U_2 + U_3 + U_4, \quad A_3 = U_1$$

Hence

$$\left. \begin{aligned} P(A_0) &= q_1 q_2 q_3, & P(A_1) &= q_1 q_2 p_3 + q_1 p_2 q_3 + p_1 q_2 q_3, \\ P(A_2) &= p_1 p_2 q_3 + p_1 q_2 p_3 + q_1 p_2 p_3, & P(A_3) &= p_1 p_2 p_3 \end{aligned} \right\} \quad (4.5.1)$$

and

$$\sum_{i=0}^3 P(A_i) = (p_1 + q_1)(p_2 + q_2)(p_3 + q_3) = 1$$

4.6 MULTINOMIAL LAW

This is another generalisation of the binomial law. Here the event space S of a random experiment E contains, instead of two points as in the Bernoullian case, m points, in general, $-U_1, U_2, \dots, U_m$ such that

$$P(U_k) = p_k \quad (k = 1, 2, \dots, m) \quad (4.6.1)$$

subject to

$$\sum_k p_k = 1 \quad (4.6.2)$$

Let a sequence of n independent trials of E be performed. The event space S^n of the compound experiment E_n will contain m^n points of the type, say, $(U_3, U_m, U_4, \dots, U_2, U_1)$. Let $A_{i_1 i_2 \dots i_m}$ denote the event ' U_1 occurs i_1 times, U_2 occurs i_2 times, ..., U_m occurs i_m times', where $\sum_k i_k = n$, which contains $\frac{n!}{i_1! i_2! \dots i_m!}$ (being the number of permutations of n things of which i_1 are alike, i_2 are alike, etc.) event points each having probability $p_1^{i_1} p_2^{i_2} \dots p_m^{i_m}$. Therefore

$$P(A_{i_1 i_2 \dots i_m}) = \frac{n!}{i_1! i_2! \dots i_m!} p_1^{i_1} p_2^{i_2} \dots p_m^{i_m} \quad (4.6.3)$$

$$(i_1, i_2, \dots, i_m = 0, 1, 2, \dots, n \text{ such that } \sum i_k = n)$$

The R.H.S. is the general term in the multinomial expansion of $(p_1 + p_2 + \dots + p_m)^n$, and hence formula (4.6.3) is called the *multinomial law*.

Now the events $A_{i_1 i_2 \dots i_m}$ ($i_1, i_2, \dots, i_m = 0, 1, 2, \dots, n$ such that $\sum i_k = n$) are pairwise mutually exclusive, and

$$\sum_{i_1 + i_2 + \dots + i_m = n} A_{i_1 i_2 \dots i_m} = S^n$$

where the summation is over all values of i_1, i_2, \dots, i_m having sum n . We then have the necessary identity :

$$\begin{aligned} P(S^n) &= \sum_{i_1 + i_2 + \dots + i_m = n} P(A_{i_1 i_2 \dots i_m}) \\ &= \sum_{i_1 + i_2 + \dots + i_m = n} \frac{n!}{i_1! i_2! \dots i_m!} p_1^{i_1} p_2^{i_2} \dots p_m^{i_m} \\ &= (p_1 + p_2 + \dots + p_m)^n = 1 \end{aligned}$$

Examples

1. A die is thrown 10 times in succession. Find the probability of the occurrence of six 4 times, five twice, and all the other faces once each.

By (4.6.3) the answer = $\frac{10!}{4! 2! 1! 1! 1! 1!} \left(\frac{1}{6}\right)^{10} \approx 0.0013$.

2. DRAWINGS WITH REPLACEMENT. An urn contains $N = N_1 + N_2 + \dots + N_m$ balls, of which N_1 are of the first colour, N_2 of the second colour, and N_m of the m th colour, and $n = i_1 + i_2 + \dots + i_m$ balls are drawn successively with replacements. Find the probability that of the balls drawn i_1 are of the first colour, i_2 of the second colour, ... and i_m of the m th colour.

Let E denote the experiment of drawing one ball from the urn and noting its colour. Then the event space S contains m points, viz. the m different colours, and the probability of the k th colour is $N_k/N = p_k$ (say), ($k = 1, 2, \dots, m$) and therefore $\sum p_k = 1$.

Now the n drawings will correspond to n independent repetitions of E , so that, by (4.6.3), the required probability

$$\begin{aligned} &= \frac{n!}{i_1! i_2! \dots i_m!} p_1^{i_1} p_2^{i_2} \dots p_m^{i_m} \\ &= \frac{n!}{i_1! i_2! \dots i_m!} \left(\frac{N_1}{N}\right)^{i_1} \left(\frac{N_2}{N}\right)^{i_2} \dots \left(\frac{N_m}{N}\right)^{i_m} \end{aligned} \quad (4.6.4)$$

DRAWINGS WITHOUT REPLACEMENT. This case has been treated in Ex. 8 Sec 3.1, and the probability is

$$\frac{\binom{N_1}{i_1} \binom{N_2}{i_2} \binom{N_m}{i_m}}{\binom{N}{n}}$$

If $N \rightarrow \infty$, subject to the condition that p_k 's are kept fixed, the limiting value of the above probability will be

$$\frac{n!}{i_1! i_2! \dots i_m!} p_1^{i_1} p_2^{i_2} \dots p_m^{i_m}$$

4.7 INFINITE SEQUENCE OF BERNOULLI TRIALS

Consider now an infinite sequence of Bernoulli trials, i.e. an infinite sequence of independent trials of a basic random experiment E whose event space S contains only two event points, viz. 'success' s and 'failure' f . Let $P(s) = p$ and $P(f) = q = 1 - p$.

The event space of the whole sequence of trials then consists of an infinite number of points such as (s, f, s, s, f, \dots) . The independence of the trials may be conveniently characterised as follows. If A_1, A_2, \dots is any infinite sequence of events each connected with the experiment E (i.e. each denotes one of the four events: success, failure, impossible event and certain event), then

$$P\{(A_1, A_2, \dots)\} = P(A_1) P(A_2) \dots \quad (4.7.1)$$

assuming that the infinite product on the right is convergent.

In particular, if in (4.7.1) $A_n = S$, the certain event for $n > r$, then we get

$$P\{(A_1, A_2, \dots, A_r, S, S, \dots)\} = P(A_1) P(A_2) \dots P(A_r) \quad (4.7.2)$$

Examples

1. In an infinite sequence of Bernoulli trials with probability of success p , find the probability that i failures will precede the first success.

In the required event we have failure f in the first i trials, success s in the $(i + 1)$ th trial and the certain event S in all subsequent trials. Hence by (4.7.2) the required probability is $q^i p$.

2. A and B toss a coin alternately and the first to obtain a head wins the toss. If A starts the game, find the probability of his winning.

Let A_n denote the event that A wins in $(2n+1)$ tosses, i.e. a head appears in the $(2n+1)$ th toss but the results of the preceding $2n$ tosses are all tails ($n=0, 1, 2, \dots$). Then by (4.7.2) $P(A_n) = (\frac{1}{2})^{2n} \cdot \frac{1}{2}$. The required event is clearly $\sum A_n$ whose probability is $P(\sum A_n) = \sum P(A_n) = \sum (\frac{1}{2})^{2n} \cdot \frac{1}{2} = \frac{2}{3}$.

4.8 MARKOV CHAINS

Let us now consider a simple but important case of dependent trials known as a *Markov chain*. Let E be a given random experiment whose event space S contains the m points U_1, U_2, \dots, U_m . Consider a sequence of n trials of E such that the outcome of any trial depends on the outcome of the immediately preceding trial but not on the outcomes of earlier trials. Let E_n denote the compound experiment of the n trials of E , the corresponding event space being S^n which contains the event points

$$(U_{i_1}, U_{i_2}, \dots, U_{i_n}) \quad (i_1, i_2, \dots, i_n = 1, 2, \dots, m)$$

Let A_i^k denote the event in space S^n that U_i occurs at the k th trial. Then

$$\begin{aligned} P\{(U_{i_1}, U_{i_2}, \dots, U_{i_n})\} &= P(A_{i_1}^1 A_{i_2}^2 \dots A_{i_n}^n) \\ &= P(A_{i_1}^1) P(A_{i_2}^2 | A_{i_1}^1) \dots P(A_{i_k}^k | A_{i_1}^1 A_{i_2}^2 \dots A_{i_{k-1}}^{k-1}) \dots \\ &\quad \dots P(A_{i_n}^n | A_{i_1}^1 A_{i_2}^2 \dots A_{i_{n-1}}^{n-1}) \end{aligned}$$

Since the outcome of the k th trial depends only on that of the $(k-1)$ th trial, we have

$$P(A_{i_k}^k | A_{i_1}^1 A_{i_2}^2 \dots A_{i_{k-1}}^{k-1}) = P(A_{i_k}^k | A_{i_{k-1}}^{k-1}) \quad (4.8.1)$$

and also assume that the conditional probability on R.H.S. is the same for all trials so that it depends on the indices i_{k-1} and i_k only, and we may write

$$P(A_j^k | A_i^{k-1}) = p_{ij} \quad (i, j = 1, 2, \dots, m) \quad (4.8.2)$$

It follows that

$$\left. \begin{aligned} p_{ij} &\geq 0 \text{ for all } i, j \\ \sum_j p_{ij} &= 1 \text{ for all } i \end{aligned} \right\} \quad (4.8.3)$$

Also let

$$P(A_i^1) = \pi_i \quad (4.8.4)$$

so that

$$\pi_i \geq 0 \text{ for all } i, \quad \sum \pi_i = 1 \quad (4.8.5)$$

$$P\{(U_{i_1}, U_{i_2}, \dots, U_{i_n})\} = \pi_i p_{i_1 i_2} p_{i_2 i_3} \dots p_{i_{n-1} i_n} \quad (4.8.6)$$

The above informal discussions lead to the following definition of a Markov chain.

A sequence of n trials of a random experiment E whose event space S contains the m event points U_1, U_2, \dots, U_m is said to be a Markov chain if numbers π_i ($i = 1, 2, \dots, m$) and p_{ij} ($i, j = 1, 2, \dots, m$) subject to (4.8.3) and (4.8.5) are given such that probabilities are assigned in S^n , the event space of the n trials of E , by (4.8.6).

Now starting from this formal definition, we may easily deduce (4.8.1), (4.8.2) and (4.8.4). Summing (4.8.6) over i_n we get

$$\begin{aligned} P(A_{i_1}^1 A_{i_2}^2 \dots A_{i_{n-1}}^{n-1}) &= P\{(U_{i_1}, U_{i_2}, \dots, U_{i_{n-1}}, S)\} \\ &= \pi_{i_1} p_{i_1 i_2} \dots p_{i_{n-2} i_{n-1}} \sum_{i_n} p_{i_{n-1} i_n} \\ &= \pi_{i_1} p_{i_1 i_2} \dots p_{i_{n-2} i_{n-1}} \quad [\text{by (4.8.3)}] \end{aligned}$$

Continuing this process of summation we have for $1 \leq k \leq n$

$$P(A_{i_1}^1 A_{i_2}^2 \dots A_{i_k}^k) = \pi_{i_1} p_{i_1 i_2} p_{i_2 i_3} \dots p_{i_{k-1} i_k} \quad (4.8.7)$$

For $k=1$, (4.8.7) reduces to (4.8.4). Since $\sum \pi_i = 1$, it follows that the sum of the numbers on the R.H.S. of (4.8.6), which are non-negative, is 1 which is a necessary condition for a valid assignment of probabilities in S^n .

Replacing k by $k-1$ in (4.8.7) and dividing (4.8.7) by this result we get

$$P(A_{i_k}^k | A_{i_1}^1 A_{i_2}^2 \dots A_{i_{k-1}}^{k-1}) = p_{i_{k-1} i_k}.$$

Also

$$P(A_{i_k}^k) = \sum \pi_{i_1} p_{i_1 i_2} \dots p_{i_{k-1} i_k} \quad (4.8.8)$$

the summation being taken over i_1, i_2, \dots, i_{k-1} , and

$$P(A_{i_{k-1}}^{k-1} A_{i_k}^k) = p_{i_{k-1} i_k} \sum \pi_{i_1} p_{i_1 i_2} \dots p_{i_{k-2} i_{k-1}}$$

the summation being over i_1, i_2, \dots, i_{k-2} .

Dividing this result by that obtained by replacing k by $k-1$ in (4.8.8) we get

$$P(A_{i_k}^k | A_{i_{k-1}}^{k-1}) = p_{i_{k-1} i_k}$$

Thus the interpretations of π_i and p_{ij} are obtained.

The $m \times m$ matrix

$$P = \begin{pmatrix} p_{11} & p_{12} \dots p_{1m} \\ p_{21} & p_{22} \dots p_{2m} \\ p_{m1} & p_{m2} \dots p_{mm} \end{pmatrix} \quad (4.8.9)$$

is called the *matrix of conditional probabilities* and the row-vector $\pi = (\pi_1, \pi_2, \dots, \pi_m)$ is said to give the *initial probability distribution* of a given Markov chain.

Note that each row sum of P is 1, and each element of P is non-negative. Likewise, the components of π are non-negative and their sum is 1.

In the theory of Markov chains it is customary to adopt the following physical analogy. Imagine a physical system which is capable of being in one of the m states U_1, U_2, \dots, U_m with different probabilities and that the system can alter its state only at times t_1, t_2, \dots, t_n (which correspond to the n trials). The conditional probability p_{ij} is then called the *probability of transition* from state U_i to state U_j , written $U_i \rightarrow U_j$, at any time, and the matrix P is called the *matrix of transition probabilities* or simply the *transition matrix* and the vector π is said to give the *initial probability distribution at time t_1* .

Higher transition probabilities. We know p_{ij} is the probability of the direct transition $U_i \rightarrow U_j$. Let us now calculate the probability of transition from U_i to U_j in r steps. This can materialise as the sequence of transitions

$$U_i \rightarrow U_{i_1} \rightarrow U_{i_2} \rightarrow \dots \rightarrow U_{i_{r-1}} \rightarrow U_j$$

where the indices i_1, i_2, \dots, i_{r-1} may vary freely.

Let the probability of transition from U_i to U_j in r steps be denoted by $p_{ij}^{(r)}$ and the matrix

$$P_r = (p_{ij}^{(r)}) \quad (i, j = 1, 2, \dots, m) \quad (4.8.10)$$

By the laws of probability

$$p_{ij}^{(r)} = \sum p_{ii_1} p_{i_1 i_2} \dots p_{i_{r-1} j} \quad (4.8.11)$$

where the summation is taken over i_1, i_2, \dots, i_{r-1} . In particular

$$p_{ij}^{(1)} = p_{ij}$$

or $P_1 = P$ and

$$p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}$$

which gives $P_2 = P^2$. Since

$$p_{ij}^{(r)} = \sum_k p_{ik} p_{kj}^{(r-1)}$$

we get $P_r = P \cdot P_{r-1}$. By induction it follows that

$$P_r = P^r \quad (4.8.12)$$

Let us now compute the (unconditional) probability that the system finds itself in state U_k at time t_r which will be denoted by $\pi_k^{(r)}$ and let the row-vector

$$\pi^{(r)} = (\pi_1^{(r)}, \pi_2^{(r)}, \dots, \pi_m^{(r)})$$

give the probability distribution at time t_r . Now the said transition may occur as transition from state U_i at time t_1 to state U_k at time t_r in $r-1$ steps where the index i may vary freely. Then

$$\pi_k^{(r)} = \sum_i \pi_i p_{ik}^{(r-1)} \quad (4.8.13)$$

or

$$\pi^{(r)} = \pi P^{r-1} \quad (4.8.14)$$

Example. Consider a Markov chain of coin tossings where the first toss is given to be a fair one and the transition probabilities of $H \rightarrow H$, $H \rightarrow T$, $T \rightarrow H$, $T \rightarrow T$ (H —head, T —tail) are respectively $\frac{1}{2}$, $\frac{1}{2}$, 0, 1. Find the probability of (i) a run of heads in first three tosses, (ii) tail in the 3rd toss assuming that there was head in the 1st toss, (iii) head in the 4th toss.

Taking $U_1 = H$, $U_2 = T$, the transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$$

and the initial distribution vector, $\pi = (\frac{1}{2}, \frac{1}{2})$.

(i) By (4.8.6) the probability of a run of heads in first 3 tosses is

$$P_1(H, H, H) = \pi_1 p_{11} p_{11} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

(ii)

$$P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ 0 & 1 \end{pmatrix}$$

The probability of tail in the 3rd toss assuming that there was head in the 1st toss is indeed the probability of transition from H to T in 2 steps, i.e. $p_{12}^{(2)} = \frac{3}{4}$.

$$(iii) \quad P^3 = \begin{pmatrix} \frac{1}{8} & \frac{7}{8} \\ 0 & 1 \end{pmatrix}$$

By (4.8.14)

$$\pi^{(4)} = \pi P^3 = \left(\frac{1}{2}, \frac{1}{2}\right) \begin{pmatrix} \frac{1}{8} & \frac{7}{8} \\ 0 & 1 \end{pmatrix} = (1/16, 15/16)$$

so that the probability of head in the 4th toss is $\pi_1^{(4)} = 1/16$.

4.9. EXERCISES

1. Three cards are successively drawn from a full pack, the card drawn being replaced every time. Find the probability of the result : spade, heart or diamond, queen. Also find the corresponding probability if the order of occurrence of the events is ignored.

2. The probability that A can solve a certain problem is $\frac{2}{3}$ and that B can solve it is $\frac{1}{3}$. If both try it independently, what is the probability that it is solved ?

3. Find the probabilities of (a) 3 heads, (b) at least 3 heads and (c) at most 3 heads in 5 throws with a coin.

4. What is the probability of obtaining multiple of three twice in a throw with 6 dice ?

5. 4 cards are drawn successively from a pack with replacements. What is the probability that all the cards are of the same suit ?

6. The probability of hitting a target is $1/5$. If 10 shots are fired, find the probability of at least two hits. Find also the minimum number of shots to be fired in order that the probability of hitting the target at least once exceeds $1/2$.

7. When a defective die is thrown 10 times the probability that an even face occurs 5 times is twice the probability that the same event occurs 4 times. Find the probability that an even face will never occur in 4 throws of the same die.

8. Suppose that the probability of a new-born baby to be a boy is $1/3$. In a family of 8 children, calculate the probability that there are 4 or 5 boys.

9. If a die is thrown n times, show that the probability of an even number of sixes is $\frac{1}{2} \{1 + (2/3)^n\}$.

10. Find the most probable number of times the event—multiple of three occurs when a die is thrown (a) 50 times, (b) 100 times.

11. Show that the most probable number of heads in $2n$ throws of a coin is n , and that the corresponding maximum probability lies between $1/2\sqrt{n}$ and $1/\sqrt{2n+1}$.

12. Prove that in n Bernoulli trials with probability of failure q , the probability of at most k successes is

$$\int_0^q x^{n-k-1}(1-x)^k dx \Big/ \int_0^1 x^{n-k-1}(1-x)^k dx$$

13. If a coin is tossed repeatedly, show that the probability of getting m heads before n tails is

$$\frac{1}{2^{m+n-1}} \sum_{i=m}^{m+n-1} \binom{m+n-1}{i}$$

14. In a Bernoullian sequence of n trials with probability of success p , find the probability that the i th success occurs at the n th trial.

15. In Banach's match-box problem, find the probability that when the first box is just emptied (i.e. the last match is drawn from the first box) the second box contains exactly i matches.

16. A class has only three students A, B, C who attend the class independently, the probabilities of their attendance on any day being $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}$ respectively. Find the probability that the total number of attendances in two consecutive days is exactly three.

17. If a die is thrown n times, find the probability that (a) the greatest, (b) the least number obtained will have a given value i .

18. From an urn containing n tickets numbered $1, 2, \dots, n$, m tickets are drawn at a time and replaced before the next drawing. Show that the probability that in k drawings each of the n tickets will appear at least once is

$$1 - \binom{n}{1} \left(\frac{n-m}{n}\right)^k + \binom{n}{2} \left(\frac{n-m}{n}\right)^k \left(\frac{n-m-1}{n-1}\right)^k - \dots$$

19. What is the probability that in a company of 500 people only one person will have birthday on New Year's day? (Assume that a year has 365 days.)

20. A card is drawn from a pack and replaced 260 times. Find the probability of obtaining queen of hearts 4 times.

21. A system consists of 1,000 connected components, where each component may fail independently of the others. If the probability that a component fails in one month is 10^{-3} , find the probability that the system will function (i.e. no component will fail) throughout a month.

22. Prove that, in a Poisson sequence of n trials, the probability of i successes is the coefficient of x^i in the product

$$(p_1x + q_1)(p_2x + q_2) \dots (p_nx + q_n)$$

where p_i denotes the probability of success in the i th trial and $q_i = 1 - p_i$.

23. Three coins having probabilities of head $1/2, 2/5, 3/7$ respectively are thrown. Find the probability of obtaining exactly one head.

24. What is the probability that the faces 1, 3, 5 turn up 2, 3, 3 times respectively in 8 throws of a die?

25. An urn contains 10 balls, of which 5 are white, 3 red and 2 black. If a ball is drawn and replaced 3 times, what is the probability that the balls are of different colours?

26. A and B alternately throw a pair of dice, A starting the game. A wins if he throws six before B throws seven, and B wins if he throws seven before A throws six. What is the probability of A 's winning?

27. Three persons toss a coin in succession and the first to obtain a head wins the game. Find their respective chances of winning.

28. A player repeatedly throws a coin and scores one point for a head and two points for a tail. If p_n denotes the probability of scoring n points, then show that $2p_n = p_{n-1} + p_{n-2}$. Hence deduce an expression for p_n , and find its limiting value as n tends to infinity.

29. If a day is dry, the conditional probability that the next day will also be dry is p ; if a day is wet, the conditional probability that the next day will be dry is p' . If u_n is the probability that the n th day will be dry, prove that

$$u_n - (p - p')u_{n-1} - p' = 0 \quad (n \geq 2)$$

If the first day is sure to be dry and $p = \frac{3}{4}$, $p' = \frac{1}{4}$, find u_n .

30. A system having three states U_1, U_2, U_3 changes its state at times $t=0, 1, 2, \dots$, the matrix of transition probabilities being

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

If it is certain that the initial state of the system is U_1 , find the probability of
(i) the event that the state of the system is U_1 at $t=0$, U_2 at $t=2$ and U_3 at $t=3$
(ii) transition from state U_3 at $t=2$ to state U_1 at $t=4$, (iii) the event that the state is U_2 at $t=4$.

31. For a two-state Markov chain with transition matrix

$$\begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}$$

and initial probability distribution (π_1, π_2) , calculate the probability distribution at the n th trial, and show that as $n \rightarrow \infty$, this distribution is independent of the initial distribution.

PROBABILITY DISTRIBUTIONS

5.1 MATHEMATICAL TOOL : FUNCTIONS ON SETS

Let us be given two sets M and N . If to every element $a \in M$ there corresponds, by some given rule, a unique element $b \in N$, then we get a *correspondence* from M to N . This correspondence is called a *function*, and we write $b = b(a)$. The set of all b 's, i.e. the function values (in the abstract sense) form a subset R of N . The set M is called the *domain of definition* of the function and the set R the *range* of the function. We at once recognise the function of a real variable to be a particular case of this, which gives a correspondence from the set of real numbers to itself.

By setting up such a correspondence, we may, as it were, pass from the set M to the set N , i.e. instead of the elements of M , we may study, if found to be more convenient, the properties of the elements of N , thereby taking an indirect account of the elements of M through the given rule of correspondence. In the theory of probability, we exactly feel such a need. We know that the results of a random experiment, i.e. the event points are in general abstract entities, and it is rather difficult to develop in detail a mathematical theory dealing with them. On the other hand, we have at our disposal the well-known theories of algebra and analysis of real numbers, and if we can pass from the abstract event space to the set of real numbers by means of a correspondence, things are expected to prove more convenient. This gives rise to the definition of what are known as random variables.

5.2 RANDOM VARIABLES

If corresponding to every point U of an event space S , we have, by a given rule, a unique real value of $X = X(U)$, i.e. X is a real-valued function defined on S , then X is called a *random* or *stochastic variable* or sometimes a *variate*. The range of the function X , i.e. the set of all values which X takes up will be called the *spectrum* of the random

variable. The spectrum may be discrete or continuous, and accordingly the random variable is said to be *discrete* or *continuous*.

Let A be a given set of real numbers. Then the set of all event points U for which $X(U) \in A$ is an event which will be denoted by $X \in A$. In particular, the event $X=a$ is the set of all event points corresponding to which X takes the value a . Similarly, we speak of the events $a < X \leq b$, $a < X < b$, $-\infty < X \leq a$, $X \neq a$ etc; the event $-\infty < X < \infty$ is obviously the certain event S so that $P(-\infty < X < \infty) = 1$.

Remarks

1. The term *random function* would have been perhaps more appropriate than random variable but is seldom used in practice.

2. HALF-OPEN INTERVALS. We know that there are three types of intervals, viz. closed, open and half-open intervals. Now the events $a \leq X \leq b$ and $b \leq X \leq c$ are not necessarily mutually exclusive but have the event $X=b$ in common; also the events $a < X < b$ and $b < X < c$ are such that their sum is not the event $a < X < c$. These difficulties are overcome if we consider half-open intervals; the events $a < X \leq b$ and $b < X \leq c$ are mutually exclusive as well as their sum is the event $a < X \leq c$. This is the reason why, in the theory of probability, we shall use the rather uncommon *half-open intervals*.

Examples

1. Consider the random experiment of throwing a coin. We may define a random variable X on its event space, which contains the two points 'head' and 'tail', by the following rule of correspondence: $X=0$ corresponding to 'tail' and $X=1$ corresponding to 'head'. Then $X=0$ and $X=1$ respectively denote the events 'tail' and 'head'. The spectrum of X consists of the two points 0 and 1, and $P(X=0) = P(X=1) = \frac{1}{2}$.

The inequality $-\infty < X < \infty$ denotes the entire event space, and we may write $(-\infty < X < \infty) = (X=0) + (X=1)$. Since $X=0$ and $X=1$ are mutually exclusive events, we must have

$$P(X=0) + P(X=1) = P(-\infty < X < \infty) = 1$$

which is true.

2. Take the event space of Ex. 5 Sec. 1.4. A random variable X may be defined by : girl $\rightarrow X=0$, boy $\rightarrow X=1$. Here $P(X=0)$ and $P(X=1)$ are not necessarily $\frac{1}{2}$ each.

3. Let the random experiment consist in throwing a die and X denote the number on the turned up face of the die. Then X is a random variable, the rule of correspondence being contained in the title itself, viz.

$$\text{one} \rightarrow X=1, \text{two} \rightarrow X=2, \dots\dots\dots \text{six} \rightarrow X=6$$

The spectrum consists of the 6 points 1, 2, ..., 6, and

$$P(X=1)=P(X=2)=\dots\dots=P(X=6)=\frac{1}{6}$$

We may define another random variable Y as follows : one, two $\rightarrow Y=0$, three, four $\rightarrow Y=1$, five, six $\rightarrow Y=2$; $Y=0$ corresponds to both the points 'one' and 'two' and hence denotes the event 'one or two'.

4. Let the experiment be drawing a card from a well-shuffled pack. If X denotes the number of points on the card drawn (assuming 11 points for the jack, 12 for the queen and 13 for the king), then X is a random variable defined on this event space of 52 points. X can take the values 1, 2, ..., 13 ; $X=1$ denotes the event 'ace' which contains 4 points, and hence $P(X=1)=1/13$ etc.

5. Consider a Bernoullian sequence of, say, 3 trials. We know that the event space S^3 contains $2^3=8$ points. Then the number of successes X is a random variable defined on S^3 , the correspondence being clearly

$$\begin{array}{ll} (F, F, F) & \rightarrow X=0 \\ \left. \begin{array}{l} (S, F, F) \\ (F, S, F) \\ (F, F, S) \end{array} \right\} & \rightarrow X=1 \\ \left. \begin{array}{l} (S, S, F) \\ (S, F, S) \\ (F, S, S) \\ (S, S, S) \end{array} \right\} & \rightarrow X=2 \\ & \rightarrow X=3 \end{array}$$

The spectrum of X consists of the 4 values 0, 1, 2, 3, and $P(X=0)=q^3$, $P(X=1)=3q^2p$, $P(X=2)=3qp^2$, $P(X=3)=p^3$. The sum of these probabilities is 1 as it must be.

6. If a ticket is drawn at random from an urn containing n tickets numbered 1, 2, ..., n and X denotes the number of the ticket drawn, then X is a random variable which can assume the values 1, 2, ..., n and

$$P(X=i)=1/n \quad (i=1, 2, \dots, n)$$

7. Let r tickets be drawn successively with replacements from an urn containing n tickets numbered $1, 2, \dots, n$. If X denotes the greatest number drawn, then the spectrum of the random variable consists of the points $1, 2, \dots, n$ and by (4.4.7)

$$P(X=i) = [i^r - (i-1)^r] / n^r \quad (i=1, 2, \dots, n)$$

8. If balls are successively drawn without replacement from an urn containing N_1 white and N_2 black balls ($N = N_1 + N_2$), then the number of black balls preceding the first white ball is a random variable which can take the values $0, 1, 2, \dots, N_2$ and by (3.1.15)

$$P(X=i) = \frac{N_1 N_2 (N_2 - 1) \dots (N_2 - i + 1)}{N(N-1) \dots (N-i)} \quad (i=1, 2, \dots, N_2)$$

$$P(X=0) = \frac{N_1}{N}$$

9. In Ex. 3 Sec. 1.4 the number of telephone calls during a fixed interval of time is a random variable, the spectrum of which is the set of all non-negative integers $0, 1, 2, \dots$.

All the random variables discussed above are of the discrete type; in Exs. 1—8 the spectrum is finite, and in Ex. 9 it is infinite.

10. In Ex. 4 Sec. 1.4 the measured value X of the length of a rod is a random variable. For a theoretical model, we may assume that X can take up any real value, i.e. X is a continuous random variable having the entire real axis as its spectrum.

5.3 DISTRIBUTION FUNCTION

The *distribution function* of a random variable X is a function of a real variable x , to be denoted by $F_x(x)$ (the subscript x pertains to the random variable X) or simply $F(x)$ defined in $(-\infty, \infty)$ by

$$F(x) = P(-\infty < X \leq x) \quad (5.3.1)$$

Basic properties of $F(x)$

Let $b > a$. The events $-\infty < X \leq a$ and $a < X \leq b$ are mutually exclusive, and

$$(-\infty < X \leq a) + (a < X \leq b) = (-\infty < X \leq b)$$

So

$$P(-\infty < X \leq a) + P(a < X \leq b) = P(-\infty < X \leq b)$$

or

$$F(a) + P(a < X \leq b) = F(b)$$

or

$$F(b) - F(a) = P(a < X \leq b) \quad (5.3.2)$$

1. By Axiom I $P(a < X \leq b) \geq 0$ so that $F(b) \geq F(a)$ for $b > a$, i.e. $F(x)$ is a monotonic non-decreasing function.

2. Let A_n denote the event $-\infty < X \leq -n$ ($n = 1, 2, \dots$). Then $\{A_n\}$ is a contracting sequence of events such that $\lim A_n = O$, the impossible event. Now

$$P(\lim A_n) = P(O) = 0$$

and

$$P(A_n) = P(-\infty < X \leq -n) = F(-n)$$

so that

$$\lim P(A_n) = F(-\infty)$$

Since by (3.1.10) $\lim P(A_n) = P(\lim A_n)$, we get

$$F(-\infty) = 0 \quad (5.3.3)$$

3. Set $A_n = (-\infty < X \leq n)$ ($n = 1, 2, \dots$) so that $\{A_n\}$ is an expanding sequence of events and $\lim A_n = (-\infty < X < \infty) = S$. Hence

$$P(\lim A_n) = P(S) = 1$$

and

$$\lim P(A_n) = \lim F(n) = F(\infty)$$

so that by (3.1.10)

$$F(\infty) = 1 \quad (5.3.4)$$

4. For any fixed point a , take $A_n = \left(a - \frac{1}{n} < X \leq a\right)$ ($n = 1, 2, \dots$) so that the sequence $\{A_n\}$ is contracting and $\lim A_n = (X = a)$. Now

$$P(\lim A_n) = P(X = a)$$

$$P(A_n) = F(a) - F\left(a - \frac{1}{n}\right)$$

and so

$$\lim P(A_n) = F(a) - F(a - 0)$$

By (3.1.10)

$$F(a) - F(a - 0) = P(X = a) \quad (5.3.5)$$

5. Define a contracting sequence of events, $\{A_n\}$ by :

$$A_n = \left(a < X \leq a + \frac{1}{n} \right) \quad (n = 1, 2, \dots).$$

Clearly, $\lim A_n = O$, the impossible event so that

$$P(\lim A_n) = P(O) = 0$$

and

$$\lim P(A_n) = \lim \left[F\left(a + \frac{1}{n}\right) - F(a) \right] = F(a+0) - F(a)$$

By (3.1.10)

$$F(a+0) = F(a) \quad (5.3.6)$$

Thus the distribution function $F(x)$ is monotonic non-decreasing such that $F(-\infty) = 0$ and $F(\infty) = 1$; it is continuous on the right at all points, but in case $P(X=a) > 0$, $F(x)$ has a jump discontinuity* on the left at $x=a$, the height of jump being equal to $P(X=a)$.

The curve $y = F(x)$ is called the *distribution curve* which is obviously confined between the lines $y=0$ and $y=1$.

Remark. Formula (5.3.2) shows that if the distribution function $F(x)$ is given, we can find the probability that X lies in any arbitrary interval, i.e. we may say that the distribution function completely determines the probability distribution and hence is of fundamental importance.

Probability mass. Consider a linear mass distribution along the x -axis, the total mass being unity. Let the distribution of mass, which may vary from point to point, be described by a function $F(x)$ which is defined to be the mass on the left of and up to the point x , so that the mass contained in any interval $a < x \leq b$ is $F(b) - F(a)$. If now $F(x)$ is interpreted as the distribution function of a probability distribution, then this quantity is $P(a < X \leq b)$, i.e. the mass in any interval may be identified with the probability that the random variable

* A function $F(x)$ is said to have a *jump discontinuity* at $x=a$ if both the limits $F(a-0)$ and $F(a+0)$ exist but are unequal ; $F(a+0) - F(a-0)$ is called the *height of jump* at that point.

lies in that interval. This conceptual mass is called *probability mass*, and often it is convenient to think in terms of this mass analogue of probability.

There are two principal types of probability distributions, viz. the discrete and continuous types which we shall now discuss.

5.4 MATHEMATICAL TOOL : STEP FUNCTIONS

Let a function $F(x)$ be defined in the interval $a \leq x \leq b$ as follows. The interval $a \leq x \leq b$ is divided into m sub-intervals by a given set of points c_0, c_1, \dots, c_m such that $a = c_0 < c_1 < \dots < c_m = b$, and

$$\begin{aligned}
 F(x) &= f_1 & c_0 \leq x < c_1 \\
 &= f_0 + f_1 & c_1 \leq x < c_2 \\
 &\dots\dots\dots & \dots\dots\dots \\
 &= f_0 + f_1 + \dots + f_{m-1} & c_{m-1} \leq x < c_m \\
 &= f_0 + f_1 + \dots + f_{m-1} + f_m & x = c_m
 \end{aligned} \tag{5.4.1}$$

where f_0, f_1, \dots, f_m are all positive constants.

The function $F(x)$ is constant in each sub-interval $c_{k-1} \leq x < c_k$ ($k = 1, 2, \dots, m$), and we say that $F(x)$ is *piecewise constant* in $a \leq x \leq b$. Moreover, $F(x)$ has a jump discontinuity at each point c_k , it being continuous on the right but discontinuous on the left, and the height of jump at c_k , $F(c_k + 0) - F(c_k - 0) = f_k$. Such a function $F(x)$ is called a *step function* in (a, b) having steps of heights f_1, f_2, \dots, f_m at the points c_1, c_2, \dots, c_m respectively. This definition may be easily extended to a step function in $(-\infty, \infty)$, an important application of which will be found in the next section.

5.5 DISCRETE DISTRIBUTIONS

If the random variable X takes up a discrete set of values $\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots$ ($\dots, x_{-2} < x_{-1} < x_0 < x_1 < x_2, \dots$) with probabilities

$$P(X = x_i) = f_i \quad (i = 0, \pm 1, \pm 2, \dots) \tag{5.5.1}$$

then the distribution function is given by the following :

In $x_i \leq x < x_{i+1}$

$$\begin{aligned}
 F(x) &= P(-\infty < X \leq x) = P\left\{ \sum_{\alpha=-\infty}^i (X=x_\alpha) \right\} \\
 &= \sum_{\alpha=-\infty}^i P(X=x_\alpha) = \sum_{\alpha=-\infty}^i f_\alpha \quad (i=0, \pm 1, \pm 2, \dots) \quad (5.5.2)
 \end{aligned}$$

That is, $F(x)$ is a step function having a step of height f_i at each point x_i of the spectrum.

Formal discussions. For a systematic theory, we shall follow the practice of defining any new concept in terms of the distribution function. Thus formally we define the distribution of a random variable X to be *discrete* or *discontinuous* if its distribution function $F(x)$ is a step function having steps of heights f_i (> 0) at the points x_i ($i=0, \pm 1, \pm 2, \dots$), i.e. given by (5.5.2). In the first place, we must see if $F(x)$ satisfies the basic properties of a distribution function. Now $F(x)$ is a monotonic non-decreasing function, continuous on the right everywhere, $F(-\infty)=0$, and further $F(\infty)=1$ if

$$\sum_{i=-\infty}^{\infty} f_i = 1 \quad (5.5.3)$$

Thus (5.5.3) imposes a necessary restriction on the constants f_i 's.

It will now be interesting to show how, in this case, we can uniquely determine the probability distribution starting from the distribution function.

1. If a is not a step point, by (5.3.5) $P(X=a) = F(a) - F(a-0) = 0$, since $F(x)$ is continuous at $x=a$.

For a step point x_i , $P(X=x_i) = F(x_i) - F(x_i-0) = f_i > 0$. This shows that the spectrum of X consists of the step points x_i 's only, and the probability mass at x_i is f_i .

2. By (5.3.2)

$$P(a < X \leq b) = \sum_{a < x_i \leq b} f_i \quad (5.5.4)$$

where the summation is extended over all values of i such that $a < x_i \leq b$, i.e. over all the step points in the interval $a < x \leq b$.

Probability diagram. Apart from the distribution curve $y = F(x)$, we may also conveniently represent a discrete distribution graphically as follows. At each point x_i of the spectrum we draw an ordinate equal to f_i , the probability mass at that point; the resulting diagram is called a *probability diagram*.

Examples

1. A random variable X can assume the values $-1, 0, 1$ with probabilities $1/3, 1/2, 1/6$ respectively. Determine the distribution.

$$\text{in } -\infty < x < -1, F(x) = P(O) = 0$$

$$\text{in } -1 \leq x < 0, F(x) = P(X = -1) = 1/3$$

$$\text{in } 0 \leq x < 1, F(x) = P\{(X = -1) + (X = 0)\} = P(X = -1) + P(X = 0) \\ = 1/3 + 1/2 = 5/6$$

$$\text{and in } 1 \leq x < \infty, F(x) = P\{(X = -1) + (X = 0) + (X = 1)\} \\ = P(X = -1) + P(X = 0) + P(X = 1) = 1/3 + 1/2 + 1/6 = 1$$

$$2. \text{ Let } F(x) = 0 \quad -\infty < x < 0$$

$$= 1/5 \quad 0 \leq x < 1$$

$$= 3/5 \quad 1 \leq x < 3$$

$$= 1 \quad 3 \leq x < \infty$$

Show that $F(x)$ is a possible distribution function, and determine the spectrum and the probability masses of the distribution.

$F(x)$ is monotonic non-decreasing everywhere, continuous on the right at every point and $F(-\infty) = 0, F(\infty) = 1$. Hence it is a possible distribution function. In fact, $F(x)$ is a step function, the step points being 0, 1, 3 which are the points of the spectrum, and

$$P(X = 0) = F(0) - F(0 - 0) = 1/5$$

$$P(X = 1) = F(1) - F(1 - 0) = 3/5 - 1/5 = 2/5$$

$$P(X = 3) = F(3) - F(3 - 0) = 1 - 3/5 = 2/5$$

We shall now introduce some of the well-known discrete distributions.

5.6 IMPORTANT DISCRETE DISTRIBUTIONS

(a) **Causal distribution.** The spectrum consists of a single point a , and

$$P(X = a) = 1 \quad (5.6.1)$$

a is a parameter of the causal distribution, i.e. for different values of a we get different causal distributions.

Example 1. If corresponding to every point of any event space we assign $X = a$, then X takes up the only value a , and hence $X = a$ denotes the certain event so that $P(X = a) = 1$. Therefore, X is causally distributed with parameter a .

(b) **Binomial distribution.** The spectrum consists of the $n+1$ points $0, 1, 2, \dots, n$, i.e.

$$\left. \begin{aligned} x_i &= i & (i=0, 1, 2, \dots, n) \\ \text{and} & \\ f_i &= \binom{n}{i} p^i (1-p)^{n-i} \end{aligned} \right\} \quad (5.6.2)$$

where n , a positive integer and p ($0 < p < 1$) are two parameters of the binomial distribution. We note that f_i 's satisfy the necessary condition (5.5.3) (cf. Sec. 4.4)

The figures below are the distribution curve and the probability diagram of the binomial distribution for $n=4$, $p=\frac{1}{2}$.

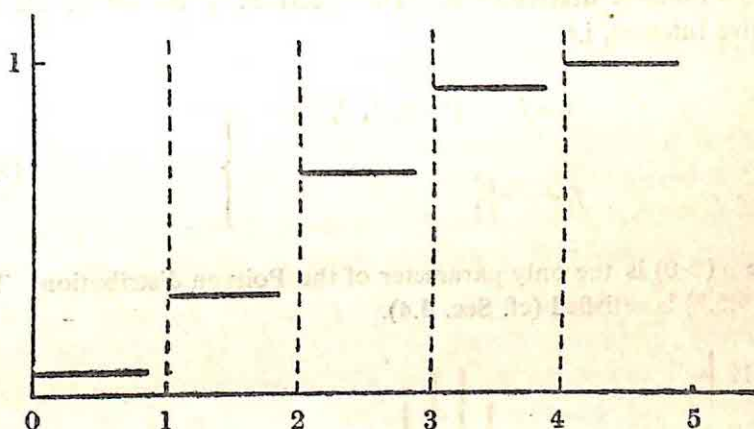


Fig. 3. Binomial Distribution Curve ($n=4$, $p=\frac{1}{2}$)

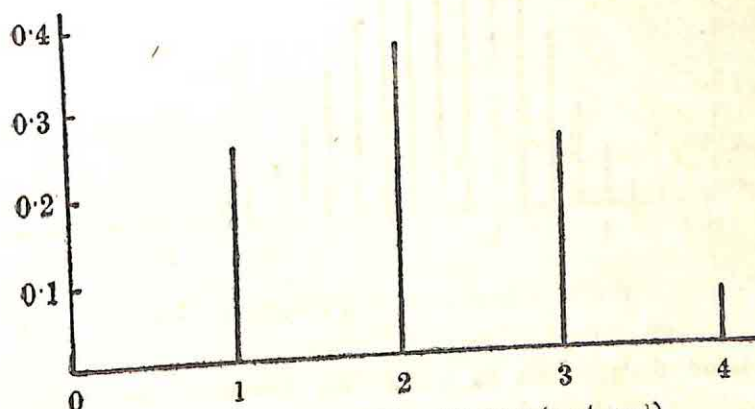


Fig. 4. Binomial Probability Diagram ($n=4$, $p=\frac{1}{2}$)

Example 2. In a Bernoullian sequence of n trials with probability of success p , the number of successes X can take the values $0, 1, 2, \dots, n$ or

$$x_i = i \quad (i = 0, 1, 2, \dots, n)$$

and by the binomial law

$$f_i = P(X = x_i) = P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

This shows that X is binomially distributed with parameters n and p .

(c) **Poisson distribution.** The spectrum is the set of all non-negative integers, i.e.

$$\left. \begin{aligned} x_i &= i \quad (i = 0, 1, 2, \dots) \\ f_i &= e^{-\mu} \frac{\mu^i}{i!} \end{aligned} \right\} \quad (5.6.3)$$

where $\mu (>0)$ is the only parameter of the Poisson distribution. Note that (5.5.3) is satisfied (cf. Sec. 4.4).

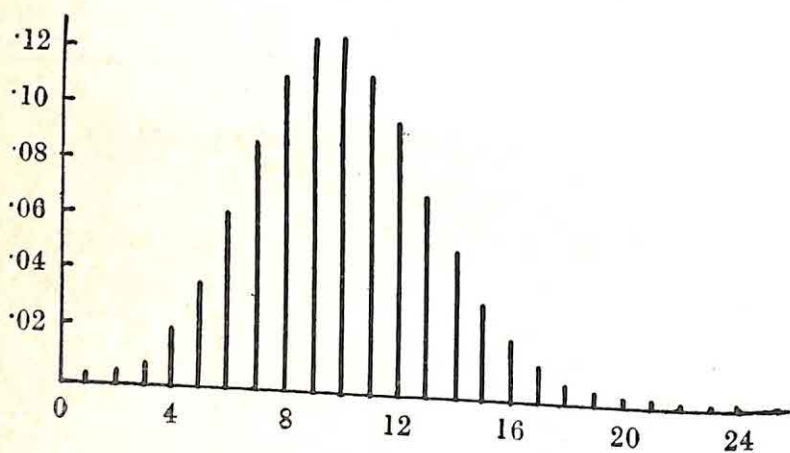


Fig. 5. Poisson Probability Diagram ($\mu = 10$)

Poisson distribution as a limiting binomial distribution. It follows from the discussions on the Poisson approximation in Sec. 4.4

that if we set $p = \mu/n$ where μ is a fixed positive number and make $n \rightarrow \infty$, we shall get in the limit

$$x_i = i \quad (i = 0, 1, 2, \dots)$$

and

$$f_i = e^{-\mu} \frac{\mu^i}{i!}$$

That is, we obtain the Poisson distribution as a limit of the binomial distribution.

Examples

3. The number of successes in a Bernoullian sequence of n trials with probability of success p , where n is very large and p very small such that $\mu = np$ is of moderate magnitude, has an approximate Poisson distribution having parameter μ .

4. There is, however, no reason to believe that the Poisson distribution can occur only as an approximation to the binomial distribution. It may also arise as itself in connection with various practical problems, and we cite the following example to illustrate this point.

POISSON PROCESS. A family of random variables $X(t)$ which depends parametrically on time t is usually called a *stochastic process*. A particular example of a stochastic process is the *Poisson process*, in which we are concerned with counting the number of changes (a general name) in a given interval of time, and which obeys the following two laws:

1. The number of changes during the interval $(t, t+h)$ is independent of the number of changes already occurred in $(0, t)$ for all t and $h (> 0)$.

2. The probability of exactly one change in $(t, t+h)$ is $\lambda h + o(h)^*$ where λ is given positive constant, and that of more than one change in the same interval is $o(h)$.

If the random variable $X(t)$ denotes the number of changes during the interval $(0, t)$, then we shall prove that $X(t)$ has a Poisson distribution.

* $o(h)$ means a function of h such that $o(h)/h \rightarrow 0$ as $h \rightarrow 0$.

Proof. * Clearly $X(t)$ can take the values $0, 1, 2, \dots$, and we write for convenience

$$P\{X(t) = i\} = P_i(t) \quad (i = 0, 1, 2, \dots)$$

Consider two successive intervals $(0, t)$ and $(t, t+h)$ ($h > 0$). Let E denote the random experiment of counting the number of changes in $(0, t)$ and E' that in $(t, t+h)$; E and E' together form the compound experiment E'' , which then consists in counting the number of changes in $(0, t+h)$. We shall interpret the first law in the sense that the random experiments E and E' are independent.

Now the event $X(t+h) = i$ (≥ 1) connected with E'' may be written as the sum of the following three pairwise mutually exclusive events :

(a) i changes in E , i.e. $X(t) = i$ and no change in E'

(b) $i-1$ changes in E , i.e. $X(t) = i-1$ and one change in E'

(c) more than one change in E' such that the total number of changes in E'' is i .

Since E and E' are independent, we have by (4.2.6) and the second law

$$P_i(t+h) = P_i(t)\{1 - \lambda h + o(h)\} + P_{i-1}(t)\{\lambda h + o(h)\} + o(h)$$

or

$$P_i(t+h) - P_i(t) = -\lambda h P_i(t) + \lambda h P_{i-1}(t) + o(h)$$

Dividing by h and passing to limit as $h \rightarrow 0$ we have

$$P_i'(t) = -\lambda P_i(t) + \lambda P_{i-1}(t) \quad (i \geq 1) \quad (i)$$

Now consider the case $i=0$. The event $X(t+h)=0$ can result in only one way, viz. no change in E , i.e. $X(t)=0$ as well as no change in E' .

Hence

$$P_0(t+h) = P_0(t)\{1 - \lambda h + o(h)\}$$

which gives

$$P_0'(t) = -\lambda P_0(t) \quad (ii)$$

We assume the obvious initial conditions

$$P_0(0) = 1 \quad (iii)$$

$$P_i(0) = 0 \quad (i \geq 1) \quad (iv)$$

Solving (ii) with the initial condition (iii) we get

$$P_0(t) = e^{-\lambda t} \quad (v)$$

Set

$$P_i(t) = e^{-\lambda t} Q_i(t) \quad (i = 0, 1, 2, \dots) \quad (vi)$$

From (v)

$$Q_0(t) = 1 \quad (vii)$$

and (iv) gives

$$Q_i(0) = 0 \quad (i \geq 1) \quad (viii)$$

Substituting (vi) in (i) we get the simple equation

$$Q_i'(t) = \lambda Q_{i-1}(t) \quad (ix)$$

We have $Q_1'(t) = \lambda Q_0(t) = \lambda$ so that using (viii)

$$Q_1(t) = \lambda t$$

Then $Q_2'(t) = \lambda Q_1(t) = \lambda^2 t$. On integration

$$Q_2(t) = \frac{(\lambda t)^2}{2}$$

Again $Q_3'(t) = \lambda Q_2(t) = \frac{(\lambda t)^2}{2}$. Hence

$$Q_3(t) = \frac{(\lambda t)^3}{3!}$$

etc. In general

$$Q_i(t) = \frac{(\lambda t)^i}{i!}$$

By (vi)

$$P_i(t) = e^{-\lambda t} \frac{(\lambda t)^i}{i!}$$

or

$$P\{X(t) = i\} = e^{-\lambda t} \frac{(\lambda t)^i}{i!} \quad (i = 0, 1, 2, \dots)$$

which shows that $X(t)$ is Poisson distributed with parameter λt .

The above laws for the Poisson process are plausibly satisfied in many cases of practical importance. For example, the number of telephone calls on a trunkline in a given interval of time, the number

of car accidents on a particular road in a given interval, the number of suicides in a particular locality in a given interval, the number of radioactive atoms disintegrating in a given sample of a radioactive substance in a given interval, the number of cosmic ray particles counted by a Geiger-Mueller counter in a given interval etc. may all be taken to be Poisson processes.

It will be proved in Chapter 7 that the parameter μ of the Poisson distribution represents what is called the average value or the mean value of the random variable. Here the average value is meant in the stochastic or probabilistic sense. Thus for a Poisson process the average number of changes in a given time interval $(0, t)$ is λt , so that the constant λ may be interpreted as the average number of changes per unit time. Anticipating this result, let us solve the following example.

Example 5. A radioactive source emits on the average 2.5 particles per second. Calculate the probability that 2 or more particles will be emitted in an interval of 4 seconds.

Here $\lambda = 2.5$ so that the number of particles emitted in an interval of 4 seconds is Poisson distributed with parameter $4\lambda = 10$. Hence the required probability is

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - P(X=0) - P(X=1) \\ &= 1 - e^{-10} - 10e^{-10} = 1 - 11e^{-10} \end{aligned}$$

Remark. In the Poisson process t may also denote any parameter other than time.

5.7 CONTINUOUS DISTRIBUTIONS

The distribution of a random variable X is said to be *continuous* if the distribution function $F(x)$ is continuous and its derivative $F'(x)$ is piecewise continuous everywhere, which means that $F(x)$ can have jump discontinuities at some points such that there are at most a finite number of them in any finite interval.

1. Since $F(x)$ is continuous at any point a , by (5.3.5)

$$P(X=a) = F(a) - F(a-0) = 0$$

2. By (5.3.2) $P(a < X \leq b) = F(b) - F(a) = \int_a^b F'(x)dx$ or

$$P(a < X \leq b) = \int_a^b f(x) dx \quad (5.7.1)$$

where

$$f(x) = F'(x) \quad (5.7.2)$$

The function $f(x)$ is called the *probability density function* of the random variable X .

3. In (5.7.1) making $a \rightarrow -\infty$ and replacing b by x , we have

$$F(x) = \int_{-\infty}^x f(x) dx \quad (5.7.3)$$

4. Since $F(x)$ is monotonic non-decreasing, it follows that

$$f(x) \geq 0 \quad \text{everywhere} \quad (5.7.4)$$

and by (5.3.4)

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (5.7.5)$$

(5.7.4) and (5.7.5) are necessary restrictions for a possible density function.

5. In differential notation we write

$$\begin{aligned} P(x < X \leq x + dx) &= F(x + dx) - F(x) = dF(x) \\ &= F'(x)dx = f(x)dx \end{aligned} \quad (5.7.6)$$

This differential is called the *probability differential* or *probability element* of X or of the corresponding distribution.

Density curve. The curve $y=f(x)$ is called the *density curve* which gives a useful graphical representation in the continuous case.

Examples

1. Find the value of the constant K such that

$$\begin{aligned} f(x) &= Kx(1-x) & 0 < x < 1 \\ &= 0 & \text{elsewhere} \end{aligned}$$

is a probability density function. Construct the distribution function and compute $P(X > \frac{1}{2})$.

By (5.7.5)

$$1 = \int_{-\infty}^{\infty} f(x)dx = K \int_0^1 x(1-x)dx = K/6 \quad \text{or} \quad K=6$$

The distribution function $F(x)$ is given by (5.7.3). In $-\infty < x < 0$, $F(x) = 0$; in $0 \leq x \leq 1$

$$F(x) = 6 \int_0^x x(1-x)dx = x^2(3-2x)$$

and in $1 < x < \infty$

$$F(x) = 6 \int_0^1 x(1-x)dx = 1$$

$$P(X > \frac{1}{2}) = \int_{\frac{1}{2}}^{\infty} f(x)dx = 6 \int_{\frac{1}{2}}^1 x(1-x)dx = \frac{1}{2}$$

The last result may also be obtained by using $F(x)$ as follows :

$$P(X > \frac{1}{2}) = 1 - P(X \leq \frac{1}{2}) = 1 - F(\frac{1}{2}) = \frac{1}{2}$$

2. Show that

$$\begin{aligned} F(x) &= 0 & -\infty < x < 0 \\ &= 1 - e^{-x} & 0 \leq x < \infty \end{aligned}$$

is a possible distribution function, and find the density function.

Since $F(x)$ is monotonic non-decreasing and continuous everywhere, $F(-\infty) = 0$ and $F(\infty) = 1$, $F(x)$ is a possible distribution function. By (5.7.2) we get on differentiation

$$\begin{aligned} f(x) &= e^{-x} & 0 < x < \infty \\ &= 0 & \text{elsewhere} \end{aligned}$$

5.8 IMPORTANT CONTINUOUS DISTRIBUTIONS

(a) **Rectangular or uniform distribution.** As the name uniform distribution implies, the density function in this case is constant in a given interval and vanishes outside it, i.e.

$$\begin{aligned} f(x) &= 0 & -\infty < x < a \\ &= \frac{1}{b-a} & a < x < b \\ &= 0 & b < x < \infty \end{aligned} \tag{5.8.1}$$

$a, b (> a)$ are the two parameters of the distribution. The necessary condition (5.7.5) is obviously satisfied. The density curve is shown in

Fig. 6, from the shape of which the other name rectangular distribution is derived.

The distribution function may be easily calculated by (5.7.3) which gives

$$\begin{aligned}
 F(x) &= 0 & -\infty < x < a \\
 &= \frac{x-a}{b-a} & a \leq x \leq b \\
 &= 1 & b < x < \infty
 \end{aligned} \tag{5.8.2}$$

The distribution curve is shown in Fig. 7.

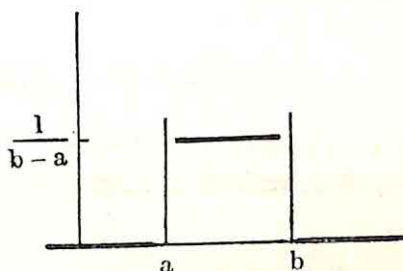


Fig. 6
Rectangular Density Curve

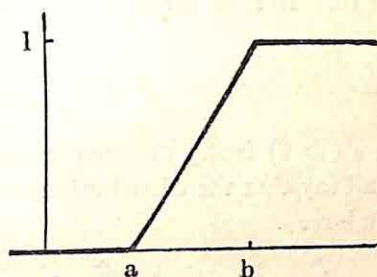


Fig. 7
Rectangular Distribution Curve

Example 1. A point X is chosen at random in the interval $a \leq x \leq b$ in such a way that the probability that it lies in any sub-interval is proportional to the length of the sub-interval. Then we can show that X is uniformly distributed over the interval (a, b) .

Proof. Let us construct the distribution function $F(x)$. From the conditions of the problem

$$\begin{aligned}
 F(x) &= 0 & -\infty < x < a \\
 &= \lambda(x-a) & a \leq x \leq b, \quad \lambda \text{ being a constant} \\
 &= 1 & b < x < \infty
 \end{aligned}$$

Since $F(b+0) = F(b)$, $\lambda = 1/(b-a)$. Hence the result.

Another method. We may also find the density function $f(x)$ by making use of the probability differential. We have

$$\begin{aligned} f(x) dx &= \lambda dx & a < x < b \\ &= 0 & \text{elsewhere} \end{aligned}$$

By (5.7.5) $\lambda = 1/(b-a)$, which proves the result.

Remark. Often, for brevity, we shall use the phrase *a point is chosen at random in a given interval* to mean that the probability of its occurrence in any sub-interval is proportional to the length of the sub-interval, i.e. the random point has a uniform distribution in the given interval.

(b) **Normal distribution.** This is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad -\infty < x < \infty \quad (5.8.3)$$

$m, \sigma (> 0)$ being the parameters of the normal distribution. We shall say that the distribution or the random-variable is normal (m, σ). We have

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-x^2} dx \\ &= \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-x^2} dx = \frac{1}{\sqrt{\pi}} \int_0^{\infty} e^{-x} x^{-\frac{1}{2}} dx = \frac{\Gamma(\frac{1}{2})}{\sqrt{\pi}} = 1 \end{aligned}$$

Hence the condition (5.7.5) is fulfilled. By (5.7.3)

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(x-m)^2}{2\sigma^2}} dx \quad (5.8.4)$$

The particular case given by $m=0, \sigma=1$ is very important and is called the *standard normal distribution*. The standard normal density and distribution functions will be denoted by the special symbols $\phi(x)$ and $\Phi(x)$ respectively which are given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx \quad (5.8.5)$$

The standard normal density and distribution curves are shown below

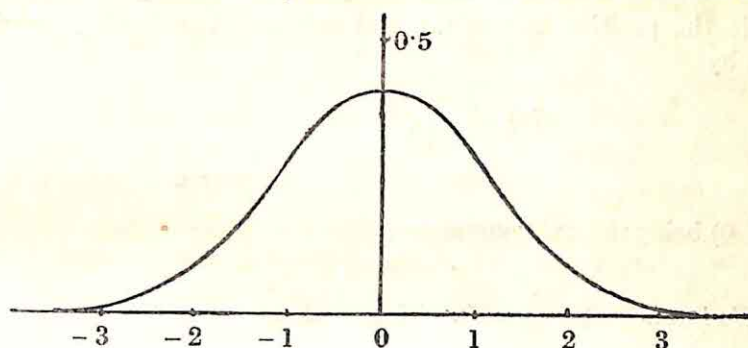


Fig. 8. Standard Normal Density Curve

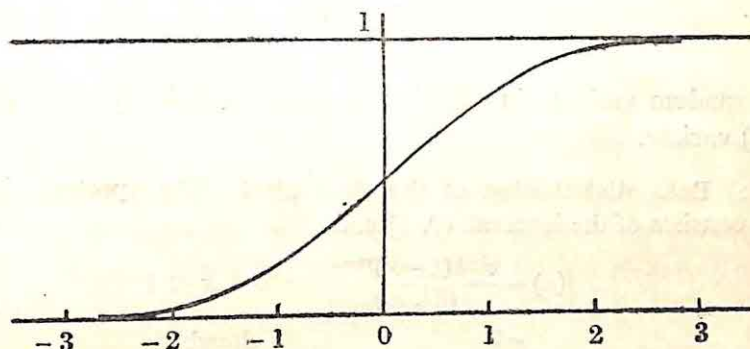


Fig. 9. Standard Normal Distribution Curve

The normal distribution is of great importance in the theory of probability, and we shall have occasion to study this distribution in detail in the course of our theory.

(c) Cauchy distribution

$$f(x) = \frac{1}{\pi} \cdot \frac{\lambda}{\lambda^2 + (x - \mu)^2} \quad -\infty < x < \infty \quad (5.8.6)$$

$\lambda (> 0)$ and μ being parameters. It may be easily verified that this density function satisfies (5.7.5). By (5.7.3)

$$F(x) = \frac{\lambda}{\pi} \int_{-\infty}^x \frac{dx}{\lambda^2 + (x - \mu)^2}$$

or

$$F(x) = \frac{1}{\pi} \tan^{-1} \left(\frac{x - \mu}{\lambda} \right) + \frac{1}{2} \quad (5.8.7)$$

(d) **Gamma distribution.** The spectrum of the gamma distribution is the positive half of the real axis, and the density function is given by

$$f(x) = \frac{e^{-x} x^{l-1}}{\Gamma(l)} \quad 0 < x < \infty \quad (5.8.8)$$

$$= 0 \quad \text{elsewhere}$$

$l (> 0)$ being the only parameter.

We know $\Gamma(l) = \int_0^{\infty} e^{-x} x^{l-1} dx$, and so

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^{\infty} f(x) dx = 1$$

The random variable in question is often conveniently referred to as a $\gamma(l)$ variate.

(e) **Beta distribution of the first kind.** The spectrum in this case consists of the interval $(0, 1)$ and

$$f(x) = \frac{x^{l-1} (1-x)^{m-1}}{B(l, m)} \quad 0 < x < 1 \quad (5.8.9)$$

$$= 0 \quad \text{elsewhere}$$

The parameters are $l (> 0)$, $m (> 0)$, and the random variable is called a $\beta_1(l, m)$ variate. Now

$$\int_{-\infty}^{\infty} f(x) dx = \frac{1}{B(l, m)} \int_0^1 x^{l-1} (1-x)^{m-1} dx = \frac{B(l, m)}{B(l, m)} = 1$$

(f) **Beta distribution of the second kind**

$$f(x) = \frac{x^{l-1}}{B(l, m)(1+x)^{l+m}} \quad 0 < x < \infty \quad (l, m > 0) \quad (5.8.10)$$

$$= 0 \quad \text{elsewhere}$$

The random variable will be called a $\beta_2(l, m)$ variate. Verify (5.7.5) remembering that we may also write

$$B(l, m) = \int_0^{\infty} \frac{x^{l-1} dx}{(1+x)^{l+m}}$$

5.9 TRANSFORMATION OF RANDOM VARIABLES

Let $y = g(x)$ be a given function of x and X a random variable defined on an event space S . Then $Y = g(X)$ is also a random variable defined on S , for corresponding to every event point of S we have a value of X , and for this value of X we get a value of $Y = g(X)$. Our problem in this section will be, given the distribution of X , to find that of Y . It can be proved in general that if $g(x)$ is any continuous function, the distribution of Y is uniquely determined by that of X . The proof of this theorem is rather difficult and beyond the scope of this book. We shall instead consider the following important particular cases.

Continuous case. Let $y = g(x)$ be a continuously differentiable function which is strictly monotonic, i.e. either $\frac{dy}{dx} > 0$ or $\frac{dy}{dx} < 0$ everywhere, so that the inverse function $x = g^{-1}(y)$ exists uniquely.

Case I. $\frac{dy}{dx} > 0$. Since $y = g(x)$ is strictly monotonic increasing, the inverse function is also strictly monotonic increasing, and hence $X \leq x$ will imply and will be implied by $g(X) \leq g(x)$ or $Y \leq y$, i.e. the events $X \leq x$ and $Y \leq y$ are identical. So $P(X \leq x) = P(Y \leq y)$ or

$$F_x(x) = F_y(y)$$

Taking differentials

$$dF_x(x) = dF_y(y) = dF \quad (\text{say})$$

or

$$dF = f_y(y)dy = f_x(x) dx = f_x(x) \frac{dx}{dy} dy$$

Hence

$$f_y(y) = f_x(x) \frac{dx}{dy} \quad (5.9.1)$$

Case II. $\frac{dy}{dx} < 0$, i.e. $g(x)$ is a strictly monotonic decreasing function. In this case

$$(X \leq x) = \{g(X) \geq g(x)\} = (Y \geq y)$$

Hence

$$P(X \leq x) = P(Y \geq y) = 1 - P(Y < y)$$

or

$$F_x(x) = 1 - F_y(y)$$

Therefore

$$dF_x(x) = -dF_y(y)$$

or

$$f_y(y)dy = -f_x(x)dx = -f_x(x) \frac{dx}{dy} dy$$

so that

$$f_y(y) = -f_x(x) \frac{dx}{dy} \quad (5.9.2)$$

The formulae (5.9.1) and (5.9.2) may be put together as

$$f_y(y) = f_x(x) \left| \frac{dx}{dy} \right| \quad (5.9.3)$$

Since in either case a unique inverse function $x = g^{-1}(y)$ exists, the R. H. S. of (5.9.3) may be expressed as a single-valued function of y .

Discrete case. The discrete case is much simpler. Here the spectrum only changes, the corresponding probability masses remaining the same.

Let $y = g(x)$ be a continuous and strictly monotonic function so that a unique inverse function $x = g^{-1}(y)$ exists. Set

$$y_i = g(x_i) \quad (5.9.4)$$

Since the transformation has a unique inverse

$$(X = x_i) = \{g(X) = g(x_i)\} = (Y = y_i)$$

Hence the spectrum of Y consists of the points y_i given by (5.9.4), and $P(X = x_i) = P(Y = y_i)$ or

$$f_{y_i} = f_{x_i} \quad (5.9.5)$$

Examples

1. The random variable X is normal (m, σ). Find the distribution of $Y = aX + b$ where a, b are constants.

Set $y = ax + b$. Hence $\frac{dy}{dx} = a$, a constant ; $x = \frac{y-b}{a}$.

As x ranges from $-\infty$ to ∞ , y also ranges over the same interval. Here

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad \left| \frac{dx}{dy} \right| = |a|$$

By (5.9.3)

$$\begin{aligned} f_y(y) &= \frac{1}{\sqrt{2\pi}|a|\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}|a|\sigma} e^{-\frac{(y-am+b)^2}{2a^2\sigma^2}} \quad (-\infty < y < \infty) \end{aligned}$$

Hence Y is normal ($am+b$, $|a|\sigma$).

In particular, the random variable $\frac{X-m}{\sigma}$ is standard normal.

2. If X is a $\beta_2(l, m)$ variate, then show that $Y=1/X$ is a $\beta_2(m, l)$ variate.

Proof. For the running variable, set $y=1/x$. Then

$$\frac{dy}{dx} = -\frac{1}{x^2} < 0$$

$$\begin{aligned} dF &= f_x(x)dx = f_x(x) \left| \frac{dx}{dy} \right| dy = \frac{x^{l-1} x^2}{B(l, m)(1+x)^{l+m}} dy \\ &= \frac{y^{m-1}}{B(m, l)(1+y)^{m+l}} dy \quad (0 < y < \infty) \end{aligned}$$

which proves the theorem.

3. If X is a standard normal variate, then prove that $Y=\frac{1}{2}X^2$ is a $\gamma(\frac{1}{2})$ variate.

Proof. Set $y=\frac{1}{2}x^2$. Then $\frac{dy}{dx}=x$, which shows that y is not a strictly monotonic function everywhere. Moreover, as x ranges from $-\infty$ to ∞ , y ranges only from 0 to ∞ traversing the interval twice in opposite directions. This presents some difficulties, and the formula (5.9.3) is not at once applicable. We may, however, solve the problem by the following special artifice.

Let $x > 0$. The event

$$\begin{aligned}(y < Y \leq y + dy) &= \{x^2 < X^2 \leq (x + dx)^2\} \\ &= (-x - dx \leq X < -x) + (x < X \leq x + dx)\end{aligned}$$

From symmetry

$$P(-x - dx \leq X < -x) = P(x < X \leq x + dx)$$

so that

$$P(y < Y < y + dy) = 2P(x < X \leq x + dx)$$

or

$$f_y(y)dy = 2f_x(x)dx = 2f_x(x) \frac{dx}{dy} dy$$

Hence

$$f_y(y) = 2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{x} = \frac{e^{-y} y^{-1/2}}{\Gamma(\frac{1}{2})} \quad (0 < y < \infty)$$

That is, Y is a $\gamma(\frac{1}{2})$ variate.

4. If X is a binomial (n, p) variate, find the distribution of the linear function $Y = aX + b$.

We know

$$x_i = i \quad (i = 0, 1, 2, \dots, n)$$

By (5.9.4) the spectrum of Y is given by

$$y_i = ai + b \quad (i = 0, 1, 2, \dots, n)$$

and by (5.9.5)

$$f_{yi} = f_{xi} = \binom{n}{i} p^i (1-p)^{n-i}$$

5.10 EXERCISES

1. If $F(x)$ denotes the distribution function of a random variable X , then show that

$$\begin{aligned}P(a < X < b) &= F(b-0) - F(a) \\ P(a \leq X \leq b) &= F(b) - F(a-0)\end{aligned}$$

2. Let $F(x)$ be a distribution function. Prove that, for any fixed $h \neq 0$, the function

$$G(x) = \frac{1}{2h} \int_{x-h}^{x+h} F(t) dt$$

is also a distribution function.

3. The spectrum of the random variable X consists of the points $1, 2, \dots, n$ and $P(X=i)$ is proportional to $1/(i+1)$. Determine the distribution function of X . Compute $P(3 < X \leq n)$ and $P(X > 5)$.

4. The distribution function $F(x)$ of a variate X is defined as follows :

$$\begin{aligned} F(x) &= A & -\infty < x < -1 \\ &= B & -1 \leq x < 0 \\ &= C & 0 \leq x < 2 \\ &= D & 2 \leq x < \infty \end{aligned}$$

where A, B, C, D are constants. Determine the values of A, B, C, D , it being given that $P(X=0)=1/6$ and $P(X > 1)=2/3$.

5. A number is chosen at random from each of the two sets $0, 1, 2, 3$ and $0, 1, 2, 3$. Find the probability distribution of the random variable denoting the sum of the numbers chosen.

6. Five balls are drawn from an urn containing 4 white and 6 black balls. Find the probability distribution of the number of white balls drawn when the balls are drawn (a) with replacements, (b) without replacements.

7. In Banach's match-box problem (Ex. 5 Sec. 4.4) find the distribution of the number of matches left in one of the boxes when the other box is just found empty.

8. Find the probability distribution of the number of failures preceding the first success in an infinite sequence of Bernoulli trials with probability of success p .

9. If X is a binomial (n, p) variate, then show that

$$P(X \leq k) = \frac{q}{1} \int_0^1 x^{n-k-1} (1-x)^k dx \bigg/ \int_0^1 x^{n-k-1} (1-x)^k dx$$

where $q=1-p$ and k is an integer such that $0 \leq k \leq n-1$.

10. If X is Poisson distributed with parameter μ , then prove that

$$P(X \leq n) = \frac{1}{n!} \int_{\mu}^{\infty} e^{-x} x^n dx$$

where n is any positive integer.

11. If there is a war every 15 years on the average, find the probability that there will be no war in 25 years.

12. There are 500 misprints in a book of 500 pages. What is the probability that a given page will contain at most 3 misprints?

13. 100 litres of water are supposed to be polluted with 10^6 bacteria. Find the probability that a sample of 1 c.c. of the same water is free from bacteria.

14. Show that a function which is $|x|$ in $(-1, 1)$ and zero elsewhere is a possible probability density function, and find the corresponding distribution function.

15. Show that a function $f(x)$ given by

$$\begin{aligned} f(x) &= x & 0 < x < 1 \\ &= k - x & 1 < x < 2 \\ &= 0 & \text{elsewhere} \end{aligned}$$

is a probability density function for a suitable value of the constant k . Calculate the probability that the random variable lies between $1/2$ and $3/2$.

16. The probability density function of a random variable X is $A \operatorname{sech} x$. Find the value of the constant A and compute $P(X < 1)$ and $P(|X| \geq 1)$.

17. Three concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet are drawn on a target board. If a shot falls within the innermost circle 3 points are scored; if it falls within the next two rings the score is respectively 2 and 1 and the score is 0 if the shot is outside the outermost circle. If the probability density of the distance of the hit from the centre of the target is

$$\frac{2}{\pi} \frac{1}{1+r^4}$$

find the probability distribution of the score.

18. A point X is chosen at random on a line segment AB whose middle point is O . Find the probability that AX , BX and AO can form the sides of a triangle.

19. A point P is chosen at random on a circle of radius a and A is a fixed point on the circle. Show that the probability that the chord AP will exceed the length of the side of an equilateral triangle inscribed in the circle is $1/3$.

20. A point P is taken at random on a line segment AB of length $2a$. Find the probability that the area of the rectangle $AP.PB$ will exceed $\frac{1}{2}a^2$.

21. A point chosen at random in a given interval divides it into two sub-intervals. Find the probability that the ratio of the length of the left sub-interval to that of the right sub-interval is less than a constant k .

22. If X is a normal (m, σ) variate, prove that

$$P(a < X < b) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)$$

and

$$P(|X-m| > a\sigma) = 2[1 - \Phi(a)]$$

where $\Phi(x)$ denotes the standard normal distribution function.

23. If X is uniformly distributed in the interval $(-1, 1)$, find the distribution of $|X|$.

24. A point is chosen at random on a semi-circle having centre at the origin and radius unity and projected on the diameter. Prove that the distance of the point of projection from the centre has probability density

$$\frac{1}{\pi \sqrt{1-x^2}} \text{ for } -1 < x < 1$$

and zero elsewhere.

25. A straight line is drawn through a fixed point (λ, μ) ($\lambda > 0$) making an angle X , which is chosen at random in the interval $(0, \pi)$, with the y -axis. Prove that the intercept on the y -axis, Y has a Cauchy distribution with parameters λ, μ .

26. The probability density of a random variable X is $2xe^{-x^2}$ for $x > 0$ and zero otherwise. Find the probability density for X^2 .

27. If X is normal $(0, 1)$, find the distribution of e^X .

28. If X is a $\gamma(l)$ variate, find the density function for \sqrt{X} .

29. If X is a $\beta_1(l, m)$ variate, then show that $X/(1-X)$ is a $\beta_2(l, m)$ variate.

30. Find the distribution of the square of a Poisson- μ variate.



TWO-DIMENSIONAL DISTRIBUTIONS

6.1 DISTRIBUTION FUNCTION IN TWO DIMENSIONS

Let X and Y be two random variables defined on the same event space S . The *joint distribution function* $F_{x, y}(x, y)$ or simply $F(x, y)$ of X and Y , or the *distribution function of the two-dimensional random variable* (X, Y) is defined by

$$F(x, y) = P(-\infty < X \leq x, -\infty < Y \leq y) \quad (6.1.1)$$

where the event $(-\infty < X \leq x, -\infty < Y \leq y)$ means the joint occurrence of the two events $-\infty < X \leq x$ and $-\infty < Y \leq y$, i.e.
 $(-\infty < X \leq x, -\infty < Y \leq y) = (-\infty < X \leq x)(-\infty < Y \leq y)$

Properties of $F(x, y)$

1. Let $a < b, c < d$. We have

$$\begin{aligned} (-\infty < X \leq a, -\infty < Y \leq c) + (a < X \leq b, -\infty < Y \leq c) \\ = (-\infty < X \leq b, -\infty < Y \leq c) \end{aligned}$$

and the events on the L.H.S. are mutually exclusive. So

$$F(a, c) + P(a < X \leq b, -\infty < Y \leq c) = F(b, c)$$

or

$$F(b, c) - F(a, c) = P(a < X \leq b, -\infty < Y \leq c) \quad (6.1.2)$$

Since the R. H. S. of (6.1.2) is non-negative

$$F(b, c) \geq F(a, c)$$

Similarly, it follows that

$$F(a, d) - F(a, c) = P(-\infty < X \leq a, c < Y \leq d) \quad (6.1.3)$$

whence

$$F(a, d) \geq F(a, c)$$

These show that $F(x, y)$ is monotonic non-decreasing in both the variables x and y .

2. Consider the half-open rectangular region :

$$a < x \leq b, c < y \leq d$$

of the xy -plane. Clearly

$$F(b, d) + F(a, c) - F(a, d) - F(b, c) \\ = P(a < X \leq b, c < Y \leq d) \quad (6.1.4)$$

3. In (6.1.1) making $x \rightarrow -\infty$ and $y \rightarrow -\infty$ we get respectively

$$F(-\infty, y) = 0, \quad F(x, -\infty) = 0 \quad (6.1.5)$$

Also making both x and $y \rightarrow \infty$ we have

$$F(\infty, \infty) = 1 \quad (6.1.6)$$

4. From (6.1.2) and (6.1.3) it follows that

$$F(a+0, c) = F(a, c), \quad F(a, c+0) = F(a, c) \quad (6.1.7)$$

5. From (6.1.4) we get the following :

$$F(b, d) + F(b-0, c) - F(b-0, d) - F(b, c) \\ = P(X=b, c < Y \leq d) \quad (6.1.8)$$

$$F(b, d) + F(a, d-0) - F(a, d) - F(a, d-0) \\ = P(a < X \leq b, Y=d) \quad (6.1.9)$$

$$F(b, d) + F(b-0, d-0) - F(b-0, d) - F(b, d-0) \\ = P(X=b, Y=d) \quad (6.1.10)$$

Marginal distributions

Given the joint distribution function $F(x, y)$ of X and Y we may easily calculate the individual distribution functions $F_x(x)$ and $F_y(y)$ of X and Y respectively as follows.

In (6.1.1) we make $y \rightarrow \infty$, and note that the event $-\infty < Y < \infty$ is the certain event S . Then

$$(-\infty < X \leq x, -\infty < Y < \infty) = (-\infty < X \leq x) S = (-\infty < X \leq x)$$

$$\text{Hence } F(x, \infty) = P(-\infty < X \leq x) \quad \text{or} \\ F_x(x) = F(x, \infty) \quad (6.1.11)$$

Similarly

$$F_y(y) = F(\infty, y) \quad (6.1.12)$$

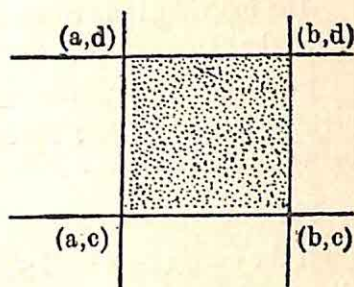


Fig. 10

The individual distributions of X and Y thus calculated are often called the *marginal distributions*.

Remark. The world marginal is obviously superfluous and perhaps serves the only purpose of hinting at the fact that they have been calculated from a joint distribution.

Independent random variables

If the events $-\infty < X \leq x$ and $-\infty < Y \leq y$ are independent for all x, y , then

$$P(-\infty < X \leq x, -\infty < Y \leq y) = P(-\infty < X \leq x) P(-\infty < Y \leq y)$$

or

$$F(x, y) = F_x(x) F_y(y) \quad (6.1.13)$$

We shall take (6.1.13) as the definition of *independence* of the random variables X and Y .

A simpler equivalent criterion for the independence of X and Y is contained in the following theorem.

Theorem I. A necessary and sufficient condition for the independence of the random variables X and Y is that their joint distribution function $F(x, y)$ can be written as the product of a function of x alone and a function of y alone.

Proof. The condition is obviously necessary. To prove its sufficiency, we note that if

$$F(x, y) = g(x)h(y)$$

then by (6.1.6)

$$g(\infty) h(\infty) = 1$$

Write

$$F(x, y) = \frac{g(x)}{g(\infty)} \frac{h(y)}{h(\infty)}$$

By (6.1.11)

$$F_x(x) = F(x, \infty) = \frac{g(x)}{g(\infty)}$$

Similarly

$$F_y(y) = \frac{h(y)}{h(\infty)}$$

Hence $F(x, y) = F_x(x) F_y(y)$, which proves the result.

Theorem II. If X and Y are independent, then

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b) P(c < Y \leq d) \quad (6.1.14)$$

Proof. By (6.1.4)

$$P(a < X \leq b, c < Y \leq d) = F(b, d) + F(a, c) - F(a, d) - F(b, c)$$

If X and Y are independent, we have using (6.1.13)

$$\begin{aligned} \text{R.H.S.} &= F_x(b)F_y(d) + F_x(a)F_y(c) - F_x(a)F_y(d) - F_x(b)F_y(c) \\ &= \{F_x(b) - F_x(a)\}\{F_y(d) - F_y(c)\} \\ &= P(a < X \leq b) P(c < Y \leq d) \end{aligned}$$

Theorem III. If X and Y are independent,

$$P(X=b, Y=d) = P(X=b) P(Y=d) \quad (6.1.15)$$

Proof. This follows immediately from (6.1.14) by making $a \rightarrow b, c \rightarrow d$, both from the left.

6.2 DISCRETE DISTRIBUTIONS

The distribution of the two-dimensional random variable (X, Y) will be called *discrete* if the distribution function $F(x, y)$ is a step function in two dimensions having steps of heights $f_{ij} (> 0)$ at the points (x_i, y_j) ($i, j = 0, \pm 1, \pm 2, \dots$), i.e.

$$F(x, y) = \sum_{\beta=-\infty}^j \sum_{\alpha=-\infty}^i f_{\alpha\beta} \quad \text{for } x_i \leq x < x_{i+1}, y_j \leq y < y_{j+1} \quad (i, j = 0, \pm 1, \pm 2, \dots) \quad (6.2.1)$$

This function will satisfy all the necessary conditions for a distribution function if

$$\sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} f_{ij} = 1 \quad (6.2.2)$$

The details of the distribution may now be obtained as follows :

1. If (b, d) is not a step point, by (6.1.10) $P(X=b, Y=d) = 0$, but at a step point (x_i, y_j)

$$\begin{aligned} P(X=x_i, Y=y_j) &= F(x_i, y_j) + F(x_i-0, y_j-0) \\ &\quad - F(x_i-0, y_j) - F(x_i, y_j-0) \end{aligned}$$

or

$$P(X=x_i, Y=y_j) = f_{ij} \quad (6.2.3)$$

That is, point probability masses f_{ij} are situated at the points (x_i, y_j) ($i, j = 0, \pm 1, \pm 2, \dots$).

2. By (6.1.4)

$$P(a < X \leq b, c < Y \leq d) = \sum_{c < y_j \leq d} \sum_{a < x_i \leq b} f_{ij} \quad (6.2.4)$$

3. The (marginal) distribution of X is given by :

For $x_i \leq x < x_{i+1}$

$$F_x(x) = F(x, \infty) = \sum_{\beta=-\infty}^{\infty} \sum_{\alpha=-\infty}^i f_{\alpha\beta} = \sum_{\alpha=-\infty}^i f_{\alpha\cdot}$$

where the symbol

$$f_{i\cdot} = \sum_{j=-\infty}^{\infty} f_{ij} \quad (6.2.5)$$

This shows that $F_x(x)$ is a step function having steps of height $f_{i\cdot}$ at x_i ($i = 0, \pm 1, \pm 2, \dots$). Hence it follows immediately from the theory of one-dimensional discrete distribution that x_i 's are indeed the points of the spectrum of X , and $P(X = x_i) = f_{i\cdot}$ or

$$f_{xi} = f_{i\cdot} \quad (6.2.6)$$

Thus if we sum all the probability masses on the line $x = x_i$ for different value of i , we get the marginal distribution of X .

Similarly, y_j 's denote the points of the spectrum of Y , and

$$f_{yj} = P(Y = y_j) = f_{\cdot j} \quad (6.2.7)$$

where

$$f_{\cdot j} = \sum_{i=-\infty}^{\infty} f_{ij} \quad (6.2.8)$$

4. The criterion of *independence* of the random variables X and Y , (6.1.13) reduces, in the discrete case, to the following :

$$f_{ij} = f_{xi} f_{yj} = f_{i\cdot} f_{\cdot j} \quad \text{for all } i, j \quad (6.2.9)$$

The necessity of the condition (6.2.9) follows at once from Theorem III Sec. 6.1.

If (6.2.9) holds, then from (6.2.1) we get :

For $x_i \leq x < x_{i+1}$, $y_j \leq y < y_{j+1}$

$$\begin{aligned} F(x, y) &= \sum_{\beta=-\infty}^j \sum_{\alpha=-\infty}^i f_{\alpha\beta} \\ &= \sum_{\beta=-\infty}^j \sum_{\alpha=-\infty}^i f_{x\alpha} f_{y\beta} = \left(\sum_{\alpha=-\infty}^i f_{x\alpha} \right) \left(\sum_{\beta=-\infty}^j f_{y\beta} \right) \end{aligned}$$

or

$$F(x, y) = F_x(x) F_y(y)$$

which holds for all x, y . Hence the condition is also sufficient.

Examples

1. An urn contains 4 white balls numbered 0,1,2,3, 3 red balls numbered 0,1,2, and 2 black balls numbered 0,1. The random experiment consists in drawing a ball at random from the urn, and two random variables X, Y are defined as follows : X takes values 0,1, and 2 respectively for white, red and black balls, and Y denotes the number of the ball. Find the joint distribution of X and Y , and deduce therefrom the marginal distributions of X, Y .

The individual spectra of X and Y are given by

$$x_i = i \quad (i = 0, 1, 2)$$

$$y_j = j \quad (j = 0, 1, 2, 3)$$

Hence the spectrum of the two-dimensional random variable (X, Y) is

$$(x_i, y_j) = (i, j) \quad (i = 0, 1, 2 ; j = 0, 1, 2, 3)$$

and $f_{ij} = P(X=i, Y=j) = \frac{1}{9}$ for all i, j , except f_{13}, f_{22}, f_{23} which are all zero. This gives the joint distribution of X and Y .

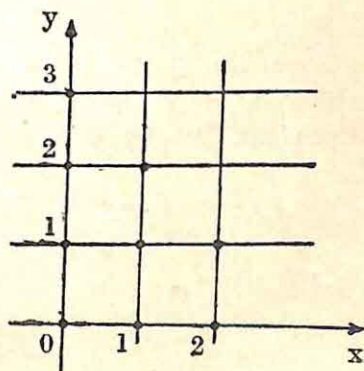


Fig. 11

Now

$$f_{xi} = f_{i\cdot} = \sum_{j=0}^3 f_{ij}$$

$$f_{x0} = \frac{4}{9}, f_{x1} = \frac{8}{9} = \frac{1}{3}, f_{x2} = \frac{2}{9}$$

$$f_{yj} = f_{\cdot j} = \sum_{i=0}^2 f_{ij}$$

$$f_{y0} = \frac{2}{9} = \frac{1}{3}, f_{y1} = \frac{8}{9} = \frac{1}{3}, f_{y2} = \frac{2}{9}, f_{y3} = \frac{1}{9}$$

These marginal distributions of X and Y may be easily verified to be correct by direct computation.

The random variables X and Y are not independent, as the condition (6.2.9) is not fulfilled.

2. Let X and Y be two Poisson variates having parameters μ_1 and μ_2 respectively. We have

$$x_i = i \quad (i=0,1,2,\dots); \quad f_{xi} = e^{-\mu_1} \frac{\mu_1^i}{i!}$$

$$y_j = j \quad (j=0,1,2,\dots); \quad f_{yj} = e^{-\mu_2} \frac{\mu_2^j}{j!}$$

Then $(x_i, y_j) = (i, j) \quad (i, j=0,1,2,\dots)$ gives the spectrum of (X, Y) . Now if X and Y are independent, by (6.2.9)

$$f_{ij} = e^{-(\mu_1 + \mu_2)} \frac{\mu_1^i}{i!} \frac{\mu_2^j}{j!}$$

Conversely, if the above expression for f_{ij} represents a joint distribution of X and Y , then we can prove that X and Y are independent. We have

$$f_{i\cdot} = e^{-(\mu_1 + \mu_2)} \frac{\mu_1^i}{i!} \sum_j \frac{\mu_2^j}{j!} = e^{-\mu_1} \frac{\mu_1^i}{i!}$$

Similarly

$$f_{\cdot j} = e^{-\mu_2} \frac{\mu_2^j}{j!}$$

Therefore $f_{ij} = f_{i\cdot} f_{\cdot j}$ for all i, j . Hence the proof.

6.3. CONTINUOUS DISTRIBUTIONS

The joint distribution of the two random variables X and Y is defined to be *continuous* if their joint distribution function $F(x, y)$ is continuous everywhere and its first and second order partial derivatives are piecewise continuous everywhere, i.e. continuous in the whole xy -plane except that there may be a finite number of curves of jump discontinuity in any bounded region.

1. Since $F(x, y)$ is continuous everywhere, by (6.1.10) the probability mass at any point (b, d) ,

$$P(X=b, Y=d)=0$$

2. We note that

$$\begin{aligned} \int_c^d \int_a^b \frac{\partial^2 F}{\partial x \partial y} dx dy &= \int_c^d \left\{ \left(\frac{\partial F}{\partial y} \right)_{(b,y)} - \left(\frac{\partial F}{\partial y} \right)_{(a,y)} \right\} dy \\ &= F(b, d) - F(b, c) - F(a, d) + F(a, c) \end{aligned}$$

By (6.1.4)

$$P(a < X \leq b, c < Y \leq d) = \int_c^d \int_a^b f(x, y) dx dy \quad (6.3.1)$$

where

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y} \quad (6.3.2)$$

$f(x, y)$ is naturally called the *joint probability density function* of X and Y .

If, instead of a rectangular region, we consider any region R of the xy -plane, then

$$P\{(X, Y) \in R\} = \iint_R f(x, y) dx dy \quad (6.3.3)$$

$$3. \quad F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(x, y) dx dy \quad (6.3.4)$$

4. Since $F(x, y)$ is monotonic in both the variables, we must have

$$f(x, y) \geq 0 \quad \text{for all } x, y \quad (6.3.5)$$

and since $F(\infty, \infty) = 1$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1 \quad (6.3.6)$$

So $f(x, y)$ must satisfy the conditions (6.3.5) and (6.3.6) in order to be a possible density function in two dimensions.

5. The probability differential

$$\begin{aligned} P(x < X \leq x + dx, y < Y \leq y + dy) \\ &= F(x + dx, y + dy) - F(x, y) - F(x + dx, y) + F(x, y + dy) \\ &= dF(x, y) \\ &= \frac{\partial^2 F}{\partial x \partial y} dx dy = f(x, y) dx dy \end{aligned} \quad (6.3.7)$$

6. The (marginal) distributions of X and Y are given by

$$F_x(x) = F(x, \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$$

Hence

$$f_x(x) = F'_x(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (6.3.8)$$

Similarly

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (6.3.9)$$

Another method. We may also easily obtain (6.3.8) and (6.3.9) by using the method of differentials. We have

$$f_x(x) dx = P(x < X \leq x + dx) = \int_{y=-\infty}^{\infty} f(x, y) dx dy = dx \int_{-\infty}^{\infty} f(x, y) dy$$

Therefore

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

Remark. The method of differentials, although possibly not very rigorous, will be sometimes found useful in our theory. In fact, it readily helps building our intuitive picture of how the probability mass is spread over the xy -plane.

7. A necessary and sufficient condition for the *independence* of X and Y , in the continuous case, is

$$f(x, y) = f_x(x) f_y(y) = \left\{ \int_{-\infty}^{\infty} f(x, y) dy \right\} \left\{ \int_{-\infty}^{\infty} f(x, y) dx \right\} \quad (6.3.10)$$

If X and Y are independent,

$$F(x, y) = F_x(x) F_y(y)$$

So

$$\frac{\partial^2 F}{\partial x \partial y} = F'_x(x) F'_y(y)$$

or

$$f(x, y) = f_x(x) f_y(y)$$

If (6.3.10) holds, then integrating we have

$$\int_{-\infty}^y \int_{-\infty}^x f(x, y) dx dy = \left\{ \int_{-\infty}^x f_x(x) dx \right\} \left\{ \int_{-\infty}^y f_y(y) dy \right\}$$

which gives (6.1.13). Hence the result.

Another method. The necessity of the condition (6.3.10) may also be proved by the method of differentials. By Theorem II Sec. 6.1, if X and Y are independent

$$\begin{aligned} P(x < X \leq x + dx, y < Y \leq y + dy) \\ = P(x < X \leq x + dx) P(y < Y \leq y + dy) \end{aligned}$$

or

$$f(x, y) dx dy = f_x(x) dx f_y(y) dy$$

or

$$f(x, y) = f_x(x) f_y(y)$$

Example. The joint probability density function of two random variables X and Y is $K(1-x-y)$ inside the triangle formed by the axes and the line $x+y=1$ and zero elsewhere. Find the value of

K and calculate $P(X < \frac{1}{2}, Y > \frac{1}{4})$. Find also the marginal distributions of X , Y , and determine whether the random variables are independent or not.

By question

$$f(x, y) = K(1 - x - y) \quad \text{for } x > 0, y > 0, x + y < 1 \\ = 0 \quad \text{elsewhere}$$

By (6.3.6)

$$1 = K \int_0^1 \int_0^{1-y} (1 - x - y) dx dy = \frac{1}{2} K \int_0^1 (1 - y)^2 dy = \frac{1}{6} K$$

so that $K = 6$. Then

$$P(X < \frac{1}{2}, Y > \frac{1}{4}) = 6 \int_0^{\frac{1}{2}} \int_{\frac{1}{4}}^{1-x} (1 - x - y) dx dy \\ = 3 \int_0^{\frac{1}{2}} (\frac{3}{4} - x)^2 dx = 13/32$$

From (6.3.8)

$$f_x(x) = 6 \int_0^{1-x} (1 - x - y) dy = 3(1 - x)^2 \quad (0 < x < 1)$$

Similarly

$$f_y(y) = 6 \int_0^{1-y} (1 - x - y) dx = 3(1 - y)^2 \quad (0 < y < 1)$$

Since $f(x, y) \neq f_x(x) f_y(y)$, X , Y are dependent.

6.4 IMPORTANT TWO-DIMENSIONAL OR BIVARIATE CONTINUOUS DISTRIBUTIONS

(a) Rectangular or uniform distribution. The density function is given by

$$f(x, y) = \frac{1}{(b-a)(d-c)} \quad \text{in } a < x < b, c < y < d \\ = 0 \quad \text{elsewhere} \quad (6.4.1)$$

The condition (6.3.6) is satisfied. There are four parameters of this distribution, viz. a, b, c, d ($b > a, d > c$).

It is easily seen that X and Y are independent having rectangular distributions in one dimension with parameters (a, b) and (c, d) respectively.

Conversely, if X and Y are independent and uniformly distributed over the intervals (a, b) and (c, d) respectively, then the two-dimensional random variable (X, Y) is uniformly distributed over the rectangle $a < x < b, c < y < d$.

We may also have a *uniform distribution* in any region R of the xy -plane, for which

$$f(x, y) = \begin{cases} \frac{1}{R} & \text{within } R \\ 0 & \text{outside } R \end{cases}$$

where R denotes the area of the region R as well.

If R' is any subregion of R , then by (6.3.3)

$$P\{(X, Y) \in R'\} = \iint_{R'} f(x, y) dx dy = \frac{R'}{R} \quad (6.4.2)$$

Let us now solve some interesting problems by the application of formula (6.4.2).

Examples

1. **BUFFON'S NEEDLE PROBLEM.** A vertical board is ruled with horizontal parallel lines at constant distance b apart. A needle of length a ($< b$) is thrown at random on the board. Find the probability that it will intersect one of the lines.

Let the random variable X denote the inclination of the needle to the horizontal and the random variable Y the perpendicular distance

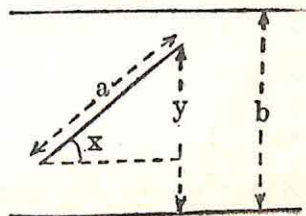


Fig. 12

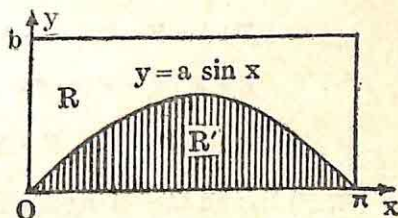


Fig. 13

of the higher end of the needle from the ruling just below it. From the conditions of the experiment we may reasonably assume X and Y to be uniformly distributed, the former over the interval $(0, \pi)$ and the latter over $(0, b)$, and that X and Y are independent. Then the two-dimensional variate (X, Y) is uniformly distributed over the region $R: 0 < x < \pi, 0 < y < b$.

Now the event that the needle intersects one of the lines can be represented by the inequalities $0 \leq Y \leq a \sin X$ or that (X, Y) lies in the region $R': 0 \leq y \leq a \sin x$. Now

$$R = \pi b, \quad R' = \int_0^{\pi} a \sin x \, dx = 2a$$

By (6.4.2) the required probability is $2a/\pi b$.

2 Two points are independently chosen at random in the interval $(-1, 1)$. Find the probability that the three parts into which the interval is divided can form the sides of a triangle.

Let the two points be represented by the random variables X and Y which are independent, each having a uniform distribution in $(-1, 1)$.

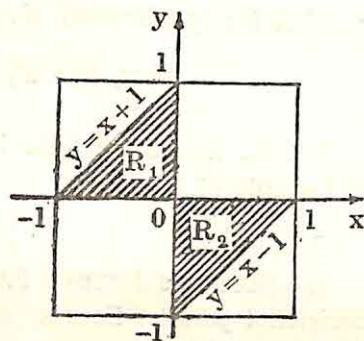


Fig. 14

If $Y > X$, the required event is represented by the following three inequalities:

$$X+1+Y-X > 1-Y \quad \text{or} \quad Y > 0$$

$$Y-X+1-Y > X+1 \quad \text{or} \quad X < 0$$

$$X+1+1-Y > Y-X \quad \text{or} \quad Y < X+1$$

i.e. (X, Y) lies in the triangular region R_1 bounded by the axes and the line $y = x + 1$.

Similarly, when $X > Y$, (X, Y) lies in the triangular region R_2 bounded by the axes and $y = x - 1$. Since

$$R = 4, R_1 = R_2 = \frac{1}{2}$$

the required probability $= \frac{R_1 + R_2}{R} = \frac{1}{4}$.

3. X and Y are independent variates, each uniformly distributed over the interval $(0, 1)$. Find the probability that the greater of X, Y is less than a fixed number k ($0 < k < 1$).

The two-dimensional random variable (X, Y) is uniformly distributed over the unit square $R: 0 < x < 1, 0 < y < 1$. The required event is $\max(X, Y) < k$.

If $Y > X$, $\max(X, Y) = Y$ and the required event means $Y < k$, or in other words, (X, Y) lies in the triangular region $R_1: x > 0, y < k, y > x$.

Similarly, in case $X > Y$, the required event is equivalent to the fact that (X, Y) lies in the triangular region $R_2: x < k, y > 0, y < x$.

Now R_1 and R_2 together form the square $R': 0 < x < k, 0 < y < k$. Hence $R = 1, R' = k^2$ and the required probability is $R'/R = k^2$.

(b) Bivariate normal distribution. Here

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right\}}$$

$$(-\infty < x < \infty, -\infty < y < \infty) \quad (6.4.3)$$

where $m_x, m_y, \sigma_x (> 0), \sigma_y (> 0)$ and $\rho (-1 < \rho < 1)$ are the five parameters of the distribution.

The (marginal) distribution of X is given by

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right\}} dy$$

Setting $x' = \frac{x - m_x}{\sigma_x}$ and $y' = \frac{y - m_y}{\sigma_y}$ we get

$$\begin{aligned} f_x(x) &= \frac{1}{2\pi\sigma_x\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{(x'^2 - 2\rho x'y' + y'^2)}{2(1-\rho^2)}} dy' \\ &= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x'^2}{2}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{(y' - \rho x')^2}{2(1-\rho^2)}} dy' \\ &= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x'^2}{2}} \end{aligned}$$

because the latter term of the product is of the form $\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(y' - m)^2}{2\sigma^2}} dy'$

and hence has value 1. Thus

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x - m_x)^2}{2\sigma_x^2}}$$

This shows that X is normal (m_x, σ_x). Similarly, Y is normal (m_y, σ_y). It is interesting to note that the individual distributions of X and Y are independent of the parameter ρ . It also follows that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = \int_{-\infty}^{\infty} f_x(x) dx = 1$$

Consider the family of ellipses given by

$$\frac{(x - m_x)^2}{\sigma_x^2} - 2\rho \frac{(x - m_x)(y - m_y)}{\sigma_x\sigma_y} + \frac{(y - m_y)^2}{\sigma_y^2} = \lambda^2 \quad (6.4.4)$$

λ being the parameter of the family. On any one of these ellipses the density function is constant, and hence these are called *equiprobability ellipses*.

On the ellipse λ

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{\lambda^2}{2(1-\rho^2)}}$$

The area of this ellipse

$$A = A(\lambda) = \frac{1}{\sqrt{\frac{1}{\lambda^4 \sigma_x^2 \sigma_y^2} - \frac{\rho^2}{\lambda^4 \sigma_x^2 \sigma_y^2}}} = \frac{\pi \lambda^2 \sigma_x \sigma_y}{\sqrt{1 - \rho^2}}$$

so that

$$dA = \frac{2\pi \lambda \sigma_x \sigma_y}{\sqrt{1 - \rho^2}} d\lambda$$

which represents the elementary area of the strip between the ellipses λ and $\lambda + d\lambda$. Hence the probability that (X, Y) lies in this strip

$$\begin{aligned} &= \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} e^{-\frac{\lambda^2}{2(1 - \rho^2)}} \cdot \frac{2\pi \lambda \sigma_x \sigma_y}{\sqrt{1 - \rho^2}} d\lambda \\ &= \frac{\lambda e^{-\frac{\lambda^2}{2(1 - \rho^2)}}}{1 - \rho^2} d\lambda \end{aligned}$$

Therefore, the probability that (X, Y) lies within the ellipse λ

$$= \frac{1}{1 - \rho^2} \int_0^\lambda \lambda e^{-\frac{\lambda^2}{2(1 - \rho^2)}} d\lambda = 1 - e^{-\frac{\lambda^2}{2(1 - \rho^2)}}$$

Hence the probability that (X, Y) lies outside the ellipse λ is $e^{-\frac{\lambda^2}{2(1 - \rho^2)}}$.

6.5 CONDITIONAL DISTRIBUTIONS

Discrete case. By definition the conditional probability

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{f_{ij}}{f_{.j}}$$

Denoting the L. H. S. by $f_{i|j}$ we have

$$f_{i|j} = \frac{f_{ij}}{f_{.j}} \quad (6.5.1)$$

We note $f_{i|j} \geq 0$ and

$$\sum_{i=-\infty}^{\infty} f_{i|j} = \frac{f_{.j}}{f_{.j}} = 1 \quad (6.5.2)$$

Hence, for a fixed value of j , $f_{i|j}$'s may denote the probability

masses of a one-dimensional discrete distribution which is called the *conditional distribution of X on the hypothesis $Y = y_j$* . Thus the conditional distribution of X on the hypothesis $Y = y_j$ is obtained by dividing every probability mass on the line $y = y_j$ by the total mass on that line.

Similarly, the conditional distribution of Y on the hypothesis $X = x_i$, is defined by

$$f_{j|i} = \frac{f_{ij}}{f_{xi}} \quad (6.5.3)$$

so that

$$\sum_{j=-\infty}^{\infty} f_{j|i} = 1 \quad (6.5.4)$$

We may write (6.5.1) and (6.5.3) in the form of a multiplication rule :

$$f_{ij} = f_{xi} f_{j|i} = f_{yj} f_{i|j} \quad (6.5.5)$$

Comparing (6.5.5) with (6.2.9) we see that the condition of independence is equivalent to

$$f_{i|j} = f_{xi} \quad \text{or} \quad f_{j|i} = f_{yj} \quad (6.5.6)$$

any one implying the other. The form (6.5.6) of the condition of independence of X and Y certainly has a ready intuitive appeal.

Example 1. Find the conditional distributions in Ex. 1 Sec. 6.2.

The conditional distribution of X on the hypothesis $Y = y_0 = 0$ is given by

$$f_{i|0} = \frac{f_{i0}}{f_{y0}} = 3f_{i0}. \quad \text{So } f_{0|0} = \frac{1}{3}, \quad f_{1|0} = \frac{1}{3}, \quad f_{2|0} = \frac{1}{3} \text{ etc.}$$

Continuous case. We have

$$\begin{aligned} P(a < X \leq b | y < Y < y + \Delta y) &= \frac{P(a < X \leq b, y < Y < y + \Delta y)}{P(y < Y < y + \Delta y)} \\ &= \frac{\int_y^{y+\Delta y} \int_a^b f(x, y) dx dy}{\int_y^{y+\Delta y} f_y(y) dy} \end{aligned}$$

We assume that $f(x, y)$ and $f_y(y)$ are continuous in their respective ranges of integration, so that using the mean-value theorem we get

$$\begin{aligned} P(a < X \leq b | y < Y < y + \Delta y) &= \frac{\Delta y \int_a^b f(x, \eta_1) dx}{\Delta y f_y(\eta_2)} \\ &\quad (y < \eta_1 = \eta_1(x), \eta_2 < y + \Delta y) \\ &= \frac{\int_a^b f(x, \eta_1) dx}{f_y(\eta_2)} \end{aligned}$$

Now making $\Delta y \rightarrow 0$ and writing

$$\lim_{\Delta y \rightarrow 0} P(a < X < b | y < Y < y + \Delta y) = P(a < X \leq b | Y = y) \quad (6.5.7)$$

we have

$$P(a < X \leq b | Y = y) = \frac{\int_a^b f(x, y) dx}{f_y(y)}$$

this step being justified by virtue of the above assumption of continuity of $f(x, y)$ and $f_y(y)$.

Setting

$$f_x(x|y) = \frac{f(x, y)}{f_y(y)} \quad (6.5.8)$$

$$P(a < X \leq b | Y = y) = \int_a^b f_x(x|y) dx \quad (6.5.9)$$

Hence $f_x(x|y) \geq 0$ and integrating (6.5.8) we get, for any fixed y

$$\int_{-\infty}^{\infty} f_x(x|y) dx = 1 \quad (6.5.10)$$

These show that, for any fixed y , $f_x(x|y)$ behaves like the density function of a one-dimensional distribution and is called the *conditional density of X on the hypothesis $Y = y$* .

The conditional distribution function $F_x(x|y)$ is given by

$$F_x(x|y) = P(-\infty < X \leq x | Y = y) = \int_{-\infty}^x f_x(x|y) dx \quad (6.5.11)$$

Similarly, we define the conditional density function of Y on the hypothesis $X=x$ by

$$f_y(y|x) = \frac{f(x, y)}{f_x(x)} \quad (6.5.12)$$

Then

$$\int_{-\infty}^{\infty} f_y(y|x) dy = 1 \quad (6.5.13)$$

and

$$F_y(y|x) = \int_{-\infty}^y f_y(y|x) dy \quad (6.5.14)$$

Combining (6.5.8) and (6.5.12) we have

$$f(x, y) = f_x(x) f_y(y|x) = f_y(y) f_x(x|y) \quad (6.5.15)$$

The criterion of independence (6.3.10) reduces to

$$f_x(x|y) = f_x(x) \quad \text{or} \quad f_y(y|x) = f_y(y) \quad (6.5.16)$$

where any one implies the other.

Remark. The conditional probability $P(a < X \leq b | Y=y)$ is as such meaningless, since the hypothesis $Y=y$, for a continuous distribution, is a stochastically impossible event. We have, however, avoided this difficulty by defining it as a limit by (6.5.7).

Another method. In terms of the differentials

$$\begin{aligned} P(x < X \leq x + dx | y < Y \leq y + dy) \\ &= \frac{P(x < X \leq x + dx, y < Y \leq y + dy)}{P(y < Y \leq y + dy)} \\ &= \frac{f(x, y) dx dy}{f_y(y) dy} = f_x(x|y) dx \end{aligned}$$

whence $f_x(x|y)$ may be interpreted as the conditional density function of X on the hypothesis $Y=y$.

Examples

2. Find the conditional probability density function $f_x(x|y)$ in the example of Sec. 6.3 and compute $P(X < \frac{1}{2} | Y = \frac{1}{4})$.

By (6.5.8)

$$f_x(x|y) = \frac{f(x, y)}{f_y(y)} = \frac{2(1-x-y)}{(1-y)^2} \quad (0 < x < 1-y)$$

where y is a fixed number such that $0 < y < 1$. If $y = \frac{1}{4}$

$$f_x(x|y) = \frac{3/2}{9/4} (\frac{3}{4} - x) \quad (0 < x < \frac{3}{4})$$

so that by (6.5.9)

$$P(X < \frac{1}{2} | Y = \frac{1}{4}) = \frac{3/2}{9/4} \int_0^{\frac{1}{2}} (\frac{3}{4} - x) dx = \frac{8}{9}$$

3. Bivariate normal distribution. Here

$$f_x(x|y) = \frac{f(x, y)}{f_y(y)} = \frac{1}{\sqrt{2\pi\sigma_x} \sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_x^2(1-\rho^2)} \left\{ x - m_x - \rho \frac{\sigma_x}{\sigma_y} (y - m_y) \right\}^2} \quad (6.5.17)$$

which shows that the conditional distribution of X on the hypothesis $Y = y$ is also normal having parameters

$$\left\{ m_x + \rho \frac{\sigma_x}{\sigma_y} (y - m_y), \sigma_x \sqrt{1 - \rho^2} \right\}$$

Similarly, the conditional distribution of Y on the hypothesis $X = x$ is normal

$$\left\{ m_y + \rho \frac{\sigma_y}{\sigma_x} (x - m_x), \sigma_y \sqrt{1 - \rho^2} \right\}$$

6.6 TRANSFORMATION OF RANDOM VARIABLES IN TWO DIMENSIONS

We shall here discuss the continuous case only, and, for simplicity, use the method of differentials. Consider a transformation of variables

$$(x, y) \rightarrow (u, v)$$

given by

$$u = u(x, y), v = v(x, y)$$

where $u(x, y)$ and $v(x, y)$ are continuously differentiable functions for which the Jacobian of the transformation, $\frac{\partial(u, v)}{\partial(x, y)}$ is either > 0 or < 0 throughout the xy -plane, so that the inverse transformation $(u, v) \rightarrow (x, y)$ is uniquely given by $x = x(u, v)$, $y = y(u, v)$.

Now given the joint density function of the two variates X and Y , we shall find that of the variates $U=u(X, Y)$ and $V=v(X, Y)$.

We note that under the above transformation the joint probability differential remains unaltered, i.e.

$$\begin{aligned} dF &= P(x < X \leq x + dx, y < Y \leq y + dy) \\ &= P(u < U \leq u + du, v < V \leq v + dv) \end{aligned}$$

or

$$f_{u,v}(u, v) du dv = f_{x,y}(x, y) dx dy = f_{x,y}(x, y) \left| \frac{\partial (x, y)}{\partial (u, v)} \right| du dv$$

giving

$$f_{u,v}(u, v) = f_{x,y}(x, y) \left| \frac{\partial (x, y)}{\partial (u, v)} \right| \quad (6.6.1)$$

the R.H.S. being expressed as a function of u, v .

Let us now prove an important theorem.

Theorem I. Let $u=u(x)$ and $v=v(y)$ be continuously differentiable and strictly monotonic functions of their respective arguments. If the random variables X and Y are independent, then so are the random variables $U=u(X)$ and $V=v(Y)$.

Proof. By (5.9.3) the distributions of U and V are given by

$$f_u(u) = f_x(x) \left| \frac{dx}{du} \right|, \quad f_v(v) = f_y(y) \left| \frac{dy}{dv} \right|$$

Let us find the joint distribution of U and V . The Jacobian of the transformation $(x, y) \rightarrow (u, v)$ reduces to $\frac{du}{dx} \frac{dv}{dy}$ which is either < 0 or > 0 everywhere, as the functions $u(x)$ and $v(y)$ are strictly monotonic.

If X and Y are independent, we have by (6.6.1)

$$f_{u,v}(u, v) = f_x(x) f_y(y) \left| \frac{dx}{du} \right| \left| \frac{dy}{dv} \right| = f_u(u) f_v(v)$$

This shows that U and V are independent.

Example 1. If the joint distribution of X and Y is the general bivariate normal distribution, and

$$U = \frac{X - m_x}{\sigma_x}, \quad V = \frac{1}{\sqrt{1 - \rho^2}} \left\{ \frac{Y - m_y}{\sigma_y} - \rho \frac{(X - m_x)}{\sigma_x} \right\}$$

then U and V are independent standard normal variates.

Proof. Set

$$u = \frac{x - m_x}{\sigma_x}, \quad v = \frac{1}{\sqrt{1 - \rho^2}} \left\{ \frac{y - m_y}{\sigma_y} - \rho \frac{(x - m_x)}{\sigma_x} \right\}$$

Then $\frac{\partial(u, v)}{\partial(x, y)} = \frac{1}{\sigma_x \sigma_y \sqrt{1 - \rho^2}}$, a positive constant. By (6.6.1)

$$\begin{aligned} f_{u,v}(u, v) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-m_x)^2}{\sigma_x^2} - 2\rho \frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2} \right\}} \\ &= \frac{1}{2\pi} e^{-(u^2+v^2)/2} = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-v^2/2} \end{aligned}$$

which proves the result.

Distribution of a function of X, Y . Let $u = u(x, y)$ be a given continuously differentiable function. To find the distribution of the variate $U = u(X, Y)$ we proceed as follows. We assume further that the function u is such that there exists another function $v = v(x, y)$ where the Jacobian $\frac{\partial(u, v)}{\partial(x, y)}$ is either $>$ or $<$ 0 everywhere. First, we find the joint distribution of U and the variate $V = v(X, Y)$ which is given by (6.6.1), and then from this joint distribution we calculate the marginal distribution of U by (6.3.8). Hence

$$f_u(u) = \int_{-\infty}^{\infty} f_{u,v}(u, v) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| dv \quad (6.6.2)$$

where the integrand is expressed as a function of u, v .

Remark. The assumption of the existence of another function $v = v(x, y)$ described above corresponds, in one dimension, to the condition that the function $y = g(x)$ is strictly monotonic (cf. Sec. 5.9).

Example 2. The joint density function of the random variables X, Y is given by

$$f(x, y) = \begin{cases} x + y & 0 < x < 1, 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the distribution of XY .

Set $U = XY, V = X$; $u = xy, v = x$ so that $x = v, y = \frac{u}{v}$ and $\frac{\partial(u, v)}{\partial(x, y)} = -x$. As x, y range from 0 to 1, u, v also range from 0 to 1. By (6.6.2)

$$f_u(u) = \int_{-\infty}^{\infty} \frac{1}{x} f(x, y) dv = \int_{-\infty}^{\infty} \frac{1}{v} f\left(v, \frac{u}{v}\right) dv$$

Now $f\left(v, \frac{u}{v}\right) = v + \frac{u}{v}$ for $0 < v < 1, 0 < \frac{u}{v} < 1$, i.e. for $u < v < 1$ ($0 < u < 1$) and zero otherwise. Hence

$$f_u(u) = \int_u^1 \left(1 + \frac{u}{v^2}\right) dv = 2(1 - u) \quad (0 < u < 1)$$

Theorem II. If X and Y are independent continuous variates, then the density function of $U = X + Y$ is given by

$$f_u(u) = \int_{-\infty}^{\infty} f_x(v) f_y(u - v) dv \quad (6.6.3)$$

Proof. Setting $u = x + y, v = x$,

$$x = v, y = u - v; \quad \frac{\partial(u, v)}{\partial(x, y)} = -1$$

Since X and Y are independent,

$$f_{x, y}(x, y) = f_x(x) f_y(y) = f_x(v) f_y(u - v)$$

(6.6.3) then follows at once from (6.6.2).

Examples

3. If X and Y are two independent normal variates (m_x, σ_x) and (m_y, σ_y) respectively, then $U = X + Y$ is a normal variate (m, σ) where $m = m_x + m_y, \sigma^2 = \sigma_x^2 + \sigma_y^2$.

Proof. By (6.6.3)

$$f_u(u) = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\{(v-m_x)^2/\sigma_x^2 + (u-v-m_y)^2/\sigma_y^2\}} dv$$

Now

$$\begin{aligned} & \frac{(v-m_x)^2}{\sigma_x^2} + \frac{(u-v-m_y)^2}{\sigma_y^2} \\ &= \frac{(v-m_x)^2}{\sigma_x^2} + \frac{(u-m-v+m_x)^2}{\sigma_y^2} \\ &= \frac{(u-m)^2}{\sigma_y^2} - \frac{2(u-m)(v-m_x)}{\sigma_y^2} + \frac{\sigma_x^2}{\sigma_x^2\sigma_y^2}(v-m_x)^2 \\ &= \frac{\sigma_x^2}{\sigma_x^2\sigma_y^2} \left\{ v-m_x - \frac{\sigma_x^2}{\sigma^2}(u-m) \right\}^2 + \frac{(u-m)^2}{\sigma^2} \end{aligned}$$

Then

$$\begin{aligned} f_u(u) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u-m)^2}{2\sigma^2}} \cdot \frac{\sigma}{\sqrt{2\pi}\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-\frac{\sigma^2}{2\sigma_x^2\sigma_y^2} \left\{ v-m_x - \frac{\sigma_x^2}{\sigma^2}(u-m) \right\}^2} dv \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(u-m)^2/2\sigma^2} \end{aligned}$$

Hence the proof.

We have thus proved a very important property of the normal distribution, viz. that the sum of two independent normal variates is again a normal variate. This is called the *reproductive property* of the normal distribution. The reproductive property is, however, not typical of the normal distribution alone but is also possessed by other distributions like the binomial, Poisson, gamma etc. Let us now prove the reproductive properties of the binomial and gamma distributions.

4. If X and Y are independent γ -variates with parameters l and m respectively, then (a) $X+Y$ is a $\gamma(l+m)$ variate and (b) X/Y is a $\beta_2(l, m)$ variate.

Proof. We can prove both (a) and (b) simultaneously if we put

$$U = X+Y, \quad V = X/Y; \quad u = x+y, \quad v = x/y$$

$$x = \frac{uv}{1+v}, \quad y = \frac{u}{1+v}; \quad \frac{\partial(u, v)}{\partial(x, y)} = -\frac{x+y}{y^2}$$

As x, y range from 0 to ∞ , u, v also range from 0 to ∞ .

Since X and Y are independent, the joint probability differential

$$\begin{aligned} dF &= \frac{e^{-x} x^{l-1}}{\Gamma(l)} \cdot \frac{e^{-y} y^{m-1}}{\Gamma(m)} dx dy = \frac{e^{-(x+y)} x^{l-1} y^{m-1}}{\Gamma(l) \Gamma(m)} \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du dv \\ &= \frac{e^{-(x+y)} x^{l-1}}{\Gamma(l) \Gamma(m)} \frac{y^{m-1}}{(x+y)} du dv = \frac{e^{-u} u^{l+m-1} v^{l-1}}{\Gamma(l) \Gamma(m) (1+v)^{l+m}} du dv \end{aligned}$$

Noting that $\Gamma(l)\Gamma(m) = B(l, m) \Gamma(l+m)$ we may write

$$dF = \frac{e^{-u} u^{l+m-1}}{\Gamma(l+m)} du \cdot \frac{v^{l-1}}{B(l, m)(1+v)^{l+m}} dv$$

which shows that U and V are independent and also proves the results (a) and (b).

5. If X and Y are independent binomial variates (n_1, p) and (n_2, p) respectively, then their sum $U = X + Y$ is a binomial $(n_1 + n_2, p)$ variate.

Proof. The spectrum of U is given by

$$u_k = k \quad (k=0, 1, 2, \dots, n_1 + n_2)$$

Then

$$\begin{aligned} f_{uk} &= P(U = u_k) = P(U = k) = \sum_{i+j=k} P(X=i, Y=j) \\ &= \sum_{i+j=k} f_{xi} f_{yj} \quad [\text{since } X \text{ and } Y \text{ are independent}] \\ &= \sum_{i+j=k} \binom{n_1}{i} \binom{n_2}{j} p^{i+j} (1-p)^{n_1+n_2-i-j} \\ &= p^k (1-p)^{n_1+n_2-k} \sum_{i+j=k} \binom{n_1}{i} \binom{n_2}{j} \end{aligned}$$

or

$$f_{uk} = \binom{n_1 + n_2}{k} p^k (1-p)^{n_1+n_2-k}$$

Hence the result.

[To show $\sum_{i+j=k} \binom{n_1}{i} \binom{n_2}{j} = \binom{n_1+n_2}{k}$ consider the identity
 $(1+x)^{n_1+n_2} = (1+x)^{n_1} (1+x)^{n_2}$

Then

$$\sum_{k=0}^{n_1+n_2} \binom{n_1+n_2}{k} x^k = \sum_{j=0}^{n_2} \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{j} x^{i+j}$$

Equating the coefficients of x^k from both sides, the above result follows.]

6.7 EXTENSIONS TO MANY DIMENSIONS. MUTUAL INDEPENDENCE

Let us first take the case of three variate X, Y, Z . The joint distribution function of X, Y and Z or the distribution function of the three-dimensional variate (X, Y, Z) will be defined by

$$F(x, y, z) = P(-\infty < X \leq x, -\infty < Y \leq y, -\infty < Z \leq z) \quad (6.7.1)$$

The (marginal) distribution of the two-dimensional variate (X, Y) is given by

$$F_{x,y}(x, y) = F(x, y, \infty) \quad (6.7.2)$$

and therefore

$$F_x(x) = F_{x,y}(x, \infty) = F(x, \infty, \infty) \quad (6.7.3)$$

and so on.

The variates (X, Y) and Z are said to be *independent* if

$$F(x, y, z) = F_{x,y}(x, y) F_z(z) \quad (6.7.4)$$

For the *continuous case*, if $f(x, y, z)$ denotes the joint density of X, Y, Z , the marginal density functions are obtained as follows.

$$f_{x,y}(x, y) = \int_{-\infty}^{\infty} f(x, y, z) dz \quad (6.7.5)$$

$$f_x(x) = \int_{-\infty}^{\infty} f_{x,y}(x, y) dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z) dy dz \quad (6.7.6)$$

etc. The condition of independence (6.7.4) becomes equivalent to

$$f(x, y, z) = f_{x,y}(x, y) f_z(z) \quad (6.7.7)$$

Theorem I. If the variates (X, Y) and Z are independent and if $u = u(x, y)$ and $w(z)$ are continuous functions of their arguments, then the variates $U = u(X, Y)$ and $W = w(Z)$ are also independent.

Proof. Let us prove the theorem for the continuous case and, for the sake of this proof, assume further that $u = u(x, y)$ and $w = w(z)$ are continuously differentiable, and the function u is such that there exists another function $v = v(x, y)$ which makes the Jacobian $\frac{\partial(u, v)}{\partial(x, y)} > \text{or} < 0$ throughout the xy -plane, and $\frac{dw}{dz} > \text{or} < 0$ everywhere. The proof will be exactly similar to that of Theorem I Sec. 6.6. We have

$$f_{u, v}(u, v) = f_{x, y}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|, \quad f_w(w) = f_z(z) \left| \frac{dz}{dw} \right|$$

Since $\frac{\partial(u, v, w)}{\partial(x, y, z)} = \frac{\partial(u, v)}{\partial(x, y)} \cdot \frac{dw}{dz} > \text{or} < 0$ everywhere, the extension of (6.6.1) for three variates gives

$$f_{u, v, w}(u, v, w) = f_{x, y, z}(x, y, z) \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right|$$

By (6.7.7)

$$f_{u, v, w}(u, v, w) = f_{x, y}(x, y) f_z(z) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| \left| \frac{dz}{dw} \right|$$

or

$$f_{u, v, w}(u, v, w) = f_{u, v}(u, v) f_w(w)$$

Integrating with respect to v from $-\infty$ to ∞ we get

$$f_{u, w}(u, w) = f_u(u) f_w(w)$$

Therefore U and W are independent.

A new concept of independence necessarily arises for more than two variates, viz. that of *mutual independence*. The variates X, Y, Z are defined to be *mutually independent* if

$$F(x, y, z) = F_x(x) F_y(y) F_z(z) \quad (6.7.8)$$

Theorem II. If X, Y, Z are mutually independent, then (a) X and Y are independent and (b) (X, Y) and Z are independent.

$$\begin{aligned} \text{Proof. } F_{x,y}(x, y) &= F(x, y, \infty) = F_x(x) F_y(y) F_z(\infty) \\ &= F_x(x) F_y(y) \end{aligned}$$

and so

$$F(x, y, z) = F_{x,y}(x, y) F_z(z)$$

These prove (a) and (b).

The generalisation of the definition of mutual independence for n variates is obvious. The n variates X_1, X_2, \dots, X_n will be called *mutually independent* if their joint distribution function

$$F(x_1, x_2, \dots, x_n) = F_{x_1}(x_1) F_{x_2}(x_2) \dots F_{x_n}(x_n) \quad (6.7.9)$$

For continuous variates, the condition (6.7.9) reduces to

$$f(x_1, x_2, \dots, x_n) = f_{x_1}(x_1) f_{x_2}(x_2) \dots f_{x_n}(x_n) \quad (6.7.10)$$

where the L.H.S. denotes the joint density function of X_1, X_2, \dots, X_n .

Theorem III. If X_1, X_2, \dots, X_n are mutually independent, then

(a) any group of $m (< n)$ of these variates are mutually independent,

(b) the variates $(X_1, \dots, X_{k_1}), (X_{k_1+1}, \dots, X_{k_2}), \dots, (X_{k_{m-1}+1}, \dots, X_n)$, where $1 < k_1 < k_2 < \dots < k_m < n$, are mutually independent, and

(c) the variates $g_1(X_1, \dots, X_{k_1}), g_2(X_{k_1+1}, \dots, X_{k_2}), \dots, g_{m+1}(X_{k_{m-1}+1}, \dots, X_n)$ where, g 's denote continuous functions of their arguments, are mutually independent.

Proof. Similar to the proofs of Theorems I and II.

Extensions of the reproductive properties

With the help of Theorem III we may now easily extend the reproductive property of any distribution to the case of n variates. For the normal distribution, in particular, we have the following general result.

If X_1, X_2, \dots, X_n are mutually independent normal variates having parameters $(m_1, \sigma_1), (m_2, \sigma_2), \dots, (m_n, \sigma_n)$ respectively, then their sum $X_1 + X_2 + \dots + X_n$ is a normal (m, σ) variate where

$$m = m_1 + m_2 + \dots + m_n, \quad \sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2$$

Proof. If X_1, X_2, \dots, X_n are mutually independent, by Theorem III(a) X_1 and X_2 are independent, and so by the reproductive property for two variates $X_1 + X_2$ is normal ($m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2}$).

Then by Theorem III(c) $X_1 + X_2$ and X_3 are independent, and hence $X_1 + X_2 + X_3$ is normal ($m_1 + m_2 + m_3, \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$).

Repeating this argument, we arrive at the general result.

Example. If three points are independently chosen at random on a circle, find the probability that they will lie on a semi-circle.

Let P, Q, R be the random points chosen on the circle. Suppose O is the centre of the circle and A a fixed point on it. Let the angles made by OP, OQ, OR with the fixed direction OA be X, Y, Z respectively. By question X, Y, Z are independent random variables, each uniformly distributed over $(0, 2\pi)$, so that the three-dimensional random variable (X, Y, Z) has a uniform distribution over the cube: $0 < x < 2\pi, 0 < y < 2\pi, 0 < z < 2\pi$, i.e. the joint density function is given by

$$f(x, y, z) = \begin{cases} \frac{1}{8\pi^3} & 0 < x < 2\pi, 0 < y < 2\pi, 0 < z < 2\pi \\ 0 & \text{elsewhere} \end{cases}$$

Consider the case $X < Y < Z$. The event in question may be

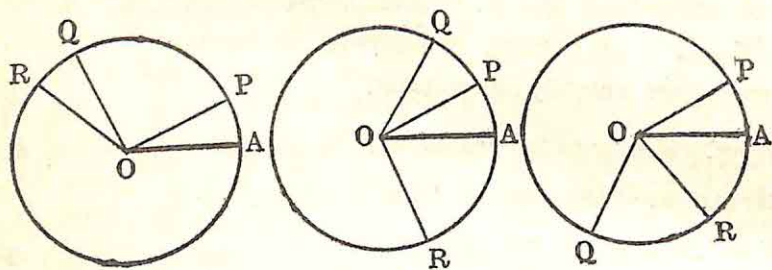


Fig. 15

expressed by the following three inequalities :

$$Z - X < \pi, Z - Y > \pi, Y - X > \pi$$

or (X, Y, Z) lies respectively in the regions

$$0 < x < 2\pi, \quad x < y < x + \pi, \quad y < z < x + \pi$$

$$0 < x < \pi, \quad x < y < \pi, \quad y + \pi < z < 2\pi$$

$$0 < x < \pi, \quad x + \pi < y < 2\pi, \quad y < z < 2\pi$$

Now

$$\begin{aligned} \int_0^{2\pi} \int_x^{x+\pi} \int_y^{y+\pi} f(x, y, z) dx dy dz &= \frac{1}{8\pi^3} \int_0^\pi \int_x^{x+\pi} \int_y^{y+\pi} dx dy dz \\ &\quad + \frac{1}{8\pi^3} \int_\pi^{2\pi} \int_x^{2\pi} \int_y^{2\pi} dx dy dz \\ &= \frac{1}{8} \cdot \frac{1}{2} + \frac{1}{8} \cdot \frac{1}{6} = \frac{1}{12} \end{aligned}$$

$$\begin{aligned} \int_0^\pi \int_x^\pi \int_{y+\pi}^{2\pi} f(x, y, z) dx dy dz &= \frac{1}{8\pi^3} \int_0^\pi \int_x^\pi \int_{y+\pi}^{2\pi} dx dy dz \\ &= \frac{1}{48} \end{aligned}$$

$$\begin{aligned} \int_0^\pi \int_{x+\pi}^{2\pi} \int_y^{2\pi} f(x, y, z) dx dy dz &= \frac{1}{8\pi^3} \int_0^\pi \int_{x+\pi}^{2\pi} \int_y^{2\pi} dx dy dz \\ &= \frac{1}{48} \end{aligned}$$

Noting that there are $3! = 6$ cases such as $X < Y < Z$ and making use of symmetry, the required probability is

$$6 \left(\frac{1}{12} + \frac{1}{48} + \frac{1}{48} \right) = \frac{3}{4}$$

6.8. EXERCISES

1. A ball is drawn from an urn containing 9 balls numbered 0, 1, 2, ..., 8, of which the first 4 are white, the next 3 red and the last 2 black. If the colours white, red and black are reckoned as colour number 0, 1 and 2 respectively, find the joint distribution of the random variables—number of the ball and the colour number. Calculate the marginal distributions of the individual random variables and the conditional distribution of the number of the ball on the assumption that the colour is white.

2. An urn contains 12 white, red and black balls, there being 4 balls of each colour numbered 0, 1, 2, 3. The random experiment consists in drawing a ball from the urn. If the colours are also assigned numbers—0 for white, 1 for red and 2 for black, then show that the number of the ball and that of the colour are independent random variables.

3. A card is drawn from a full pack of 52 cards. If X denotes the number on the card (assuming that the jack, queen and king respectively correspond to the numbers 11, 12 and 13) and Y takes values 1, 2, 3, 4 for spade, heart, diamond and club respectively, find the distribution of the two-dimensional variate (X, Y) , and show that X and Y are independent.

4. Find the joint distribution of two independent variates, one of which is Poisson distributed with parameter μ and the other binomially distributed with parameters (n, p) .

5. Show that the function $f(x, y)$ defined by

$$f(x, y) = \sin x \sin y \quad 0 < x < \frac{1}{2}\pi, \quad 0 < y < \frac{1}{2}\pi \\ = 0 \quad \text{elsewhere}$$

is a possible two-dimensional probability density function. Find the marginal density functions, and prove that the random variables are independent.

6. Determine the value of the constant K which makes

$$f(x, y) = Kxy \quad (0 < x < 1, \quad 0 < y < x)$$

a joint probability density function. Calculate the marginal density functions and show that the variates are dependent.

7. The joint probability density function of two variates X, Y is given by

$$f(x, y) = (6 - x - y)/8 \quad 0 < x < 2, \quad 2 < y < 4 \\ = 0 \quad \text{elsewhere}$$

Calculate the following probabilities :

$$P(X < 1, Y < 3), \quad P(X + Y < 3)$$

$$P(X < 1 | Y = 3), \quad P(X < 1 | Y < 3)$$

8. If $f(x, y) = 3x^2 - 8xy + 6y^2$ ($0 < x < 1, 0 < y < 1$), find $f_x(x|y)$ and $f_y(y|x)$, and show that X and Y are dependent.

9. The joint density function of the random variables X, Y is given by :

$$f(x, y) = 2 \quad (0 < x < 1, \quad 0 < y < x)$$

Find the marginal and conditional density functions. Compute $P(\frac{1}{4} < X < \frac{3}{4} | Y = \frac{1}{2})$.

10. Let X, Y be two random variables, each having spectrum $(-\infty, \infty)$. If the conditional density function of X on the hypothesis $Y = y$ is $|y|e^{-x^2 y^2} / \sqrt{\pi}$ and the density function of Y is $\lambda e^{-\lambda^2 x^2} / \sqrt{\pi}$, prove that the density function of X is $\lambda / \pi (x^2 + \lambda^2)$.

11. Two points are independently chosen at random in the interval $(0, 1)$. Find the probability that the distance between them is less than a fixed number k ($0 < k < 1$).

12. Two numbers are independently chosen at random between 0 and 1. Show that the probability that their product is less than a constant k ($0 < k < 1$) is $k(1 - \log k)$.

13. If p and q are independent variates each uniformly distributed over the interval $(-1, 1)$, find the probability that the equation $x^2 + 2px + q = 0$ has real roots.

14. A dart is thrown at random on a square target board having vertices $(1, 0)$, $(0, 1)$, $(-1, 0)$ and $(0, -1)$, the point at which the dart hits the board being (X, Y) . Find the marginal density functions of X and Y and show that they are dependent.

15. A random point (X, Y) is uniformly distributed over a circular region : $x^2 + y^2 < a^2$. Find the marginal distributions of X and Y , and the conditional distribution of Y assuming that $X = x$ ($|x| < a$).

16. A straight line AB is divided by a point C into two parts AC and CB whose lengths are a and b respectively. If two points P and Q are independently chosen at random on AC and CB respectively, find the probability that AP , PQ , QB can form the sides of a triangle.

17. A floor is paved with tiles, each tile being a parallelogram. If a stick of length c falls on the floor parallel to a diagonal whose length is l , then show that the probability that it will lie entirely on one tile is $(1 - c/l)^2$. Show also that if the distances between pairs of opposite sides of a tile are a and b and a circle of diameter d falls on the floor, the probability that it will lie on one tile is $(1 - d/a)(1 - d/b)$.

18. Two points P, Q are independently chosen at random on a circle and A is a fixed point also on the circle. Find the probability that the three points A, P, Q will lie on the same semi-circle.

19. In the bivariate normal distribution the equiprobability ellipse λ for which the probability mass in the strip between the ellipses λ and $\lambda + d\lambda$ is maximum (for fixed $d\lambda$) is called the *ellipse of maximum probability*. Find the probability mass outside this ellipse of maximum probability.

20. Let (X, Y) have the general two-dimensional normal distribution, and we make a linear transformation :

$$U = (X - m_x) \cos \alpha + (Y - m_y) \sin \alpha, \quad V = -(X - m_x) \sin \alpha + (Y - m_y) \cos \alpha$$

Show that U, V will be independent normal variates if

$$\tan 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

21. If $f(x, y) = x + y$ ($0 < x < 1, 0 < y < 1$) is the joint density function of the variates X and Y , find the distribution of $X + Y$.

22. If X and Y are independent variates both uniformly distributed over $(0, 1)$, find the distributions of $X + Y$, $X - Y$ and XY .

23. If the Cartesian co-ordinates of a random point are independent standard normal variates, show that its polar co-ordinates are also independent variates, and find their distributions.

24. If X_1, X_2 are independent random variables each having the density function $2xe^{-x^2}$ ($0 < x < \infty$), find the density function for the random variable $\sqrt{X_1^2 + X_2^2}$.

25. Let X_1, X_2 be independent variates each having the density function ae^{-ax} ($0 < x < \infty$), where a is a positive constant. Find the density function for X_1/X_2 . Prove that the variate $X_2/(X_1 + X_2)$ is uniformly distributed over $(0, 1)$.

26. If X, Y are independent random variables whose density functions are given by

$$f_x(x) = \frac{1}{\pi\sqrt{1-x^2}} \quad (-1 < x < 1), \quad f_y(y) = 2ye^{-y^2} \quad (0 < y < \infty)$$

prove that the random variable XY has a normal distribution.

27. Prove that the sum of two independent Poisson variates having parameters μ_1 and μ_2 is a Poisson variate having parameter $\mu_1 + \mu_2$.

28. If X and Y are independent Cauchy variates having parameters (λ_1, μ_1) and (λ_2, μ_2) respectively, then show that $X+Y$ also has a Cauchy distribution having parameters $(\lambda_1 + \lambda_2, \mu_1 + \mu_2)$. Hence deduce that if X_1, X_2, \dots, X_n are n mutually independent Cauchy variates each with parameters (λ, μ) , their arithmetic mean $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$ is a Cauchy variate with the same parameters (λ, μ) .

29. If the joint probability density function of the random variables X, Y, Z is given by

$$f(x, y, z) = 6/(1+x+y+z)^4 \quad (0 < x < \infty, 0 < y < \infty, 0 < z < \infty)$$

find the distribution of the random variable $X+Y+Z$.

30. If X_1, X_2, \dots, X_n are mutually independent random variables each uniformly distributed over $(0, 1)$, prove that the density function of $X = X_1 X_2 \dots X_n$ is

$$\frac{1}{(n-1)!} \left[\log \left(\frac{1}{x} \right) \right]^{n-1} \quad (0 < x < 1)$$

Hence deduce the density function of the geometric mean $(X_1 X_2 \dots X_n)^{1/n}$.

31. The numbers X_1, X_2, \dots, X_n are independently chosen at random in the interval (a, b) . Prove that the probability density function of the random variable $X = \min(X_1, X_2, \dots, X_n)$ is given by

$$f(x) = \frac{n(b-x)^{n-1}}{(b-a)^n} \quad (a < x < b)$$

MATHEMATICAL EXPECTATIONS I

7.1 MATHEMATICAL EXPECTATION OR MEAN VALUE

It is sometimes worthwhile to reckon the salient features of a mass distribution in terms of a number of typical values. For example, if we know the centre of gravity and the moments of inertia about any three mutually perpendicular lines of a mass distribution in space, we certainly get some rough idea regarding the distribution. In the theory of probability also, we shall be interested in obtaining a number of such typical values which will be called the *characteristics* of the distribution. For this, we start with the definition of what is called mathematical expectation or mean value.

If $g(x)$ is a continuous function, then we know that the distribution of the random variable $g(X)$ is completely determined by that of X , and we define the *mathematical expectation* or the *mean value* of the function $g(X)$ of the random variable X , to be denoted by $E\{g(X)\}$, by

$$E\{g(X)\} = \sum_{i=-\infty}^{\infty} g(x_i) f_i \quad \text{for a discrete distribution} \quad (7.1.1)$$

$$= \int_{-\infty}^{\infty} g(x) f(x) dx \quad \text{for a continuous distribution}$$

provided the series or the infinite integral *converges absolutely*. This implies that if the series or integral in question is not absolutely convergent (even if it is convergent but not absolutely so), we shall say that the expectation does not exist. We note that $E\{g(X)\}$ is a constant associated with the distribution of X and the function $g(x)$.

Some simple properties

The following simple properties of mathematical expectation may be easily verified.

1. $E(a) = a$, a being a constant.

2. $E\{ag(X)\} = aE\{g(X)\}$ ($a = \text{constant}$).
3. $E\{g_1(X) + g_2(X) + \cdots + g_n(X)\}$
 $= E\{g_1(X)\} + E\{g_2(X)\} + \cdots + E\{g_n(X)\}$
4. $|E\{g(X)\}| \leq E\{|g(X)|\}$
5. If $g(x) \geq 0$ everywhere, then $E\{g(X)\} \geq 0$.
6. If $g(x) \geq 0$ everywhere and $E\{g(X)\} = 0$, then $g(X) = 0$, i.e. the random variable $g(X)$ has a one-point distribution at $x = 0$.

Remark. If we set $Y = g(X)$, we get two definitions of the mathematical expectation of this random variable—the first to be calculated from the distribution of X which is represented by $E\{g(X)\}$ and the second $E(Y)$ which is calculated from the distribution of Y . Now in order that our definition of mathematical expectation is unambiguous, these two numbers must be the same, i.e. $E\{g(X)\} = E(Y)$. This result can indeed be proved to be true for any continuous function $g(x)$. Let us here prove it for a continuous distribution under the simplifying assumption that $g(x)$ is a continuously differentiable and strictly monotonic, say, increasing function. We have

$$E(Y) = \int_{-\infty}^{\infty} y f_y(y) dy$$

Putting $y = g(x)$ and using (5.9.1)

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} g(x) f_x(x) \frac{dx}{dy} \frac{dy}{dx} dx \\ &= \int_{-\infty}^{\infty} g(x) f_x(x) dx = E\{g(X)\} \end{aligned}$$

Examples

1. From an urn containing N_1 white and N_2 black balls ($N = N_1 + N_2$), balls are successively drawn without replacement. Find the mathematical expectation of the number of black balls preceding the first white ball.

Let X denote the number of black balls preceding the first white ball. Then the random variable X can take the values $0, 1, 2, \dots, N_2$, i.e. $x_i = i$ ($i = 0, 1, 2, \dots, N_2$), and the corresponding probability masses $f_i = P(X=i)$ are given in Ex. 8 Sec. 5.2. Hence

$$E(X) = \sum_1^{N_2} i \frac{N_1 N_2 (N_2 - 1) \dots (N_2 - i + 1)}{N(N-1) \dots (N-i)}$$

Since $\sum f_i = 1$, we get

$$\sum_1^{N_2} \frac{N_2 (N_2 - 1) \dots (N_2 - i + 1)}{(N-1) \dots (N-i)} = \frac{N_2}{N_1}$$

which is an identity in N_1, N_2 . Replacing N_1 by $N_1 + 1$ in this identity we get another identity :

$$\sum_1^N \frac{N_2 (N_2 - 1) \dots (N_2 - i + 1)}{N(N-1) \dots (N-i+1)} = \frac{N_2}{N_1 + 1}$$

Taking the difference of these two identities we have

$$E(X) = \frac{N_2}{N_1 + 1}$$

2. If r tickets are drawn successively with replacements from an urn containing n tickets numbered $1, 2, \dots, n$, then find the expectation of the greatest number drawn.

The distribution of X , the greatest number drawn, is given in Ex. 7 Sec. 5.2. This gives

$$\begin{aligned} E(X) &= n^{-r} \sum_1^n i [i^r - (i-1)^r] \\ &= n^{-r} \sum_1^n [i^{r+1} - (i-1)^{r+1} - (i-1)^r] \\ &= n - n^{-r} \sum_1^n (i-1)^r \end{aligned}$$

If n is large

$$n^{-r} \sum_1^n (i-1)^r \simeq n \int_0^1 x^r dx = \frac{n}{r+1}$$

so that

$$E(X) \simeq \frac{nr}{r+1}$$

3. A point is chosen at random on a circle of radius a . Compute the mathematical expectation of its distance from a fixed point also on the circle.

Let O be the centre of the circle, A the fixed point and P the random point on the circle. Let X denote the angle POA . By question, the random variable X is uniformly distributed over the interval $(0, 2\pi)$, i.e. the probability density function is given by

$$f(x) = 1/2\pi \quad (0 < x < 2\pi)$$

Now $PA = 2a \sin \frac{1}{2}X$, and its expectation is

$$E(2a \sin \frac{1}{2}X) = \frac{a}{\pi} \int_0^{2\pi} \sin \frac{1}{2}x \, dx = \frac{4a}{\pi}$$

7.2 MEAN

The *mean* of X or *that of the corresponding distribution* is naturally defined to be $E(X)$ and is often denoted by the special symbol $m(X)$ or m_x or simply m , i.e.

$$m = E(X) \quad (7.2.1)$$

The mean has an important physical significance, viz. it represents the centre of mass of the probability mass distribution. The mean thus gives a rough position of the bulk of the distribution and, as such, is called a *measure of location*. The mean, however, is not the only measure of location; there are other measures of location as well, some of which will be defined later.

Examples

1. BINOMIAL DISTRIBUTION

$$\begin{aligned} m &= \sum_{i=0}^n i \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^n i \binom{n}{i} p^i (1-p)^{n-i} \\ &= np \sum_{i=1}^n \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} = np \sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-1-i} \\ &= np \end{aligned}$$

2. POISSON DISTRIBUTION

$$m = \sum_{i=0}^{\infty} i e^{-\mu} \frac{\mu^i}{i!} = e^{-\mu} \sum_{i=1}^{\infty} \frac{\mu^i}{(i-1)!} = \mu e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i}{i!} = \mu e^{-\mu} e^{\mu} = \mu$$

3. NORMAL DISTRIBUTION

$$\begin{aligned} m(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-(x-m)^2/2\sigma^2} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-m) e^{-(x-m)^2/2\sigma^2} dx \\ &\quad + \frac{m}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-m)^2/2\sigma^2} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-x^2/2\sigma^2} dx + m = 0 + m = m \end{aligned}$$

Moreover, the integral $\int_{-\infty}^{\infty} x e^{-x^2/2\sigma^2} dx$, and hence the integral representing $m(X)$, is absolutely convergent. Therefore, the mean exists and is equal to m . Thus we see that the parameter m has the natural interpretation of being the mean of the distribution.

4. CAUCHY DISTRIBUTION. This is an example in which the mean does not exist.

$$m = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{x dx}{\lambda^2 + (x-\mu)^2} = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{(x-\mu) dx}{\lambda^2 + (x-\mu)^2} + \mu = \frac{\lambda}{\pi} \int_{-\infty}^{\infty} \frac{x dx}{\lambda^2 + x^2} + \mu$$

Since the last integral is not absolutely convergent, the mean does not exist. (The integral is also not convergent, but its Cauchy principal value exists which is zero.)

5. GAMMA DISTRIBUTION

$$m = \frac{1}{\Gamma(l)} \int_0^{\infty} x e^{-x} x^{l-1} dx = \frac{1}{\Gamma(l)} \int_0^{\infty} e^{-x} x^l dx = \frac{\Gamma(l+1)}{\Gamma(l)} = l$$

7.3 MOMENTS

Let k be a non-negative integer. The *moment of order k* or the *k th moment of X about a fixed point a* is defined to be the mean value $E\{(X-a)^k\}$.

$E\{|X-a|^k\}$ will be called the *k th absolute moment of X about a* . It is to be noted that the existence of the k th moment implies the existence of the k th absolute moment. Again if the k th moment exists, we can prove that the $(k-1)$ th moment also exists. This follows easily from the theories of convergence of series and infinite integrals, in view of the inequality $|x-a|^{k-1} \leq |x-a|^k + 1$ for all x . Hence if the k th moment exists, all moments of order less than k exist.

The k th moment about the origin, which is often simply called the *k th moment of X or its distribution*, will be denoted by $\alpha_k(X)$ or α_{xk} or α_k , i.e

$$\alpha_k = E(X^k) \quad (7.3.1)$$

Clearly $\alpha_0 = 1$, $\alpha_1 = m$, i.e the mean is the first moment of the distribution.

Of special interest are the moments about the mean which are also called the *central moments*. The *k th central moment $\mu_k(X)$ or μ_{xk} or μ_k* is then given by

$$\mu_k = E\{(X-m)^k\} \quad (7.3.2)$$

We have $\mu_0 = 1$, $\mu_1 = 0$ for all random variables.

The central moments μ_k may be expressed in terms of the ordinary moments of order $\leq k$ as follows.

$$(X-m)^k = \sum_{i=0}^k (-1)^i \binom{k}{i} X^{k-i} m^i$$

which gives

$$\mu_k = \sum_{i=0}^k (-1)^i \binom{k}{i} \alpha_{k-i} m^i \quad (7.3.3)$$

Noting $\alpha_0 = 1$ and $\alpha_1 = m$, we get

$$\begin{aligned}\mu_2 &= \alpha_2 - m^2 \\ \mu_3 &= \alpha_3 - 3\alpha_2 m + 2m^3 \\ \mu_4 &= \alpha_4 - 4\alpha_3 m + 6\alpha_2 m^2 - 3m^4 \\ &\dots \quad \dots \quad \dots\end{aligned}\tag{7.3.4}$$

We can calculate the moments of a continuous function $g(X)$ of X from the distribution of X as follows.

Setting $Y = g(X)$, $\alpha_k(Y) = E[\{g(X)\}^k]$ or

$$\alpha_k\{g(X)\} = E[\{g(X)\}^k]\tag{7.3.5}$$

For the central moments, we note

$$m_y = E(Y) = E\{g(X)\}\tag{7.3.6}$$

Hence

$$\mu_k(Y) = E\{(Y - m_y)^k\} = E[\{g(X) - m_y\}^k]$$

or

$$\mu_k\{g(X)\} = E[\{g(X) - m_y\}^k]\tag{7.3.7}$$

where m_y is given by (7.3.6).

If $Y = aX + b$, a, b being constants, $m_y = am_x + b$. So

$$\mu_k(aX + b) = E\{(aX + b - am_x - b)^k\} = a^k E\{(X - m_x)^k\}$$

or

$$\mu_k(aX + b) = a^k \mu_k(X)\tag{7.3.8}$$

Examples

1. NORMAL DISTRIBUTION. Since the mean is m ,

$$\begin{aligned}\mu_k &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - m)^k e^{-(x-m)^2 / 2\sigma^2} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - m)^{k-1} (x - m) e^{-(x-m)^2 / 2\sigma^2} dx\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sigma} \left\{ -\sigma^2 (x-m)^{k-1} e^{-(x-m)^2/2\sigma^2} \right. \\
&\quad \left. + (k-1)\sigma^2 \int_{-\infty}^{\infty} (x-m)^{k-2} e^{-(x-m)^2/2\sigma^2} dx \right\} \\
&= (k-1)\sigma^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-m)^{k-2} e^{-(x-m)^2/2\sigma^2} dx
\end{aligned}$$

or

$$\mu_k = (k-1)\sigma^2 \mu_{k-2}$$

Since $\mu_0 = 1$, $\mu_1 = 0$, it follows that

$$\mu_{2k+1} = 0, \quad \mu_{2k} = 1 \cdot 3 \cdot 5 \dots (2k-1) \sigma^{2k}$$

2. GAMMA DISTRIBUTION

$$\begin{aligned}
a_k &= \frac{1}{\Gamma(l)} \int_0^{\infty} x^k e^{-x} x^{l-1} dx = \frac{1}{\Gamma(l)} \int_0^{\infty} e^{-x} x^{l+k-1} dx = \frac{\Gamma(l+k)}{\Gamma(l)} \\
&= l(l+1)(l+2) \dots (l+k-1)
\end{aligned}$$

7.4 VARIANCE

The second central moment μ_2 is of great importance and is called the *variance of X* written as $\text{var}(X)$, i.e.

$$\text{var}(X) = \mu_2 = E\{(X-m)^2\} \quad (7.4.1)$$

It is clear from the definition that the variance is a characteristic which describes how widely the probability masses are spread about the mean or, in other words, an inverse measure of concentration of the probability masses about the mean. Such a measure will be called a *measure of dispersion*. The physical interpretation of the variance is that it represents the moment of inertia of the probability mass distribution about a line through the mean perpendicular to the line of the distribution.

Now the variance is a quantity whose unit is square of the unit of the random variable, and it is sometimes more convenient to have a measure of dispersion having the same unit as that of the random

variable. This is obtained by taking the positive square root of the variance which is called the *standard deviation* of X to be denoted by $\sigma(X)$ or σ_x or σ , i.e.

$$\sigma = + \sqrt{\text{var}(X)} \quad (7.4.2)$$

1. Since $(x-m)^2 \geq 0$ for all x , $\text{var}(X) = 0$ implies $X = m$, i.e. the whole mass is concentrated at the mean.

2. The second moment about any point is minimum when taken about the mean.

$$\begin{aligned} \text{Proof. } (X-a)^2 &= \{(X-m) + (m-a)\}^2 \\ &= (X-m)^2 + 2(m-a)(X-m) + (m-a)^2 \end{aligned}$$

So

$$\begin{aligned} E\{(X-a)^2\} &= E\{X-m\}^2 + 2(m-a)E(X-m) + (m-a)^2 \\ &= \mu_2 + (m-a)^2 \geq \mu_2 \end{aligned}$$

3. The following formulae will be sometimes useful for evaluating the variance :

$$\sigma^2 = \alpha_2 - m^2 \quad (7.4.3)$$

$$\sigma_2 = E\{X(X-1)\} - m(m-1) \quad (7.4.4)$$

Proof. (7.4.3) is nothing but the first equation of the set (7.3.4). For (7.4.4) we note

$$(X-m)^2 = X(X-1) - 2mX + X + m^2$$

Hence

$$\begin{aligned} \sigma^2 &= E\{X(X-1)\} - 2mE(X) + E(X) + m^2 \\ &= E\{X(X-1)\} - 2m^2 + m + m^2 = \text{R.H.S. of (7.4.4)} \end{aligned}$$

4. For $k=2$, (7.3.8) gives

$$\text{var}(aX+b) = a^2 \text{var}(X) \quad (7.4.5)$$

Putting $a=0$, it follows that the variance of a constant is zero.

In terms of standard deviations (7.4.5) may be written in the form :

$$\sigma(aX+b) = |a| \sigma(X) \quad (7.4.6)$$

Standardised or normalised random variable. For any random variable X , if we set

$$X^* = \frac{X - m}{\sigma} \quad (7.4.7)$$

then $m(X^*) = 0$, $\sigma(X^*) = 1$, and X^* is called the *standardised* or *normalised random variable corresponding to X* . The standardised random variable, we note, is dimensionless, and as such all its moments are also so, and since its mean is zero, its central moments coincide with the corresponding ordinary moments.

It follows that the standardised variable corresponding to a linear function $aX + b$ is $\frac{a}{|a|} X^*$, which reduces simply to X if $a > 0$.

Examples

1. **BINOMIAL DISTRIBUTION.** In this case the formula (7.4.4) will be convenient, and we have

$$\begin{aligned} E\{X(X-1)\} &= \sum_{i=0}^n i(i-1) \binom{n}{i} p^i (1-p)^{n-i} \\ &= n(n-1)p^2 \sum_{i=2}^n \binom{n-2}{i-2} p^{i-2} (1-p)^{n-i} = n(n-1)p^2 \end{aligned}$$

From Ex. 1 Sec. 7.2 $m = np$, so that

$$\sigma^2 = n(n-1)p^2 - np(np-1) = np(1-p)$$

2. POISSON DISTRIBUTION

$$E\{X(X-1)\} = \sum_{i=0}^{\infty} i(i-1) e^{-\mu} \frac{\mu^i}{i!} = e^{-\mu} \mu^2 \sum_{i=2}^{\infty} \frac{\mu^{i-2}}{(i-2)!} = \mu^2$$

Hence

$$\sigma^2 = \mu^2 - \mu(\mu-1) = \mu, \quad \sigma = \sqrt{\mu}$$

3. **NORMAL DISTRIBUTION.** In Ex. 1 Sec. 7.3 we deduced the general formula for μ_{2k} from which $\mu_2 = \sigma^2$. Hence the parameter σ of the normal distribution denotes nothing but its standard deviation.

4. **GAMMA DISTRIBUTION.** By Ex. 2 Sec. 7.3 $\alpha_2 = l(l+1)$. Therefore by (7.4.3)

$$\sigma^2 = l(l+1) - l^2 = l, \quad \sigma = \sqrt{l}$$

5. CAUCHY DISTRIBUTION. Since the mean of this distribution is non-existent, the variance necessarily does not exist. We note further that $E\{(X-\mu)^2\}$, the second moment of X about μ , is infinitely large.

7.5 THIRD CENTRAL MOMENT

It is sometimes necessary to study the degree of lack of symmetry of a distribution, and the third central moment provides a measure for this *asymmetry* or *skewness*. We note that, for a symmetrical distribution, the point about which the distribution is symmetrical naturally becomes the mean, and all central moments of odd order are zero, i.e. $\mu_{2k+1} = 0$ ($k = 0, 1, 2, \dots$). The first central moment μ_1 , however, vanishes for all distributions. Hence we may take μ_3 to be a measure of skewness. Now it is desirable to have a dimensionless measure describing such a property as asymmetry or skewness. This is obtained by considering the third moment of the standardised variate X^* , viz μ_3/σ^3 , and we define the *coefficient of skewness* γ_1 by

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \quad (7.5.1)$$

When $\gamma_1 > 0$ the density curve for a continuous variate may be roughly described as having a longer 'tail' on the positive or the right side than on the negative or the left side, and conversely for negative skewness. The figure below (Fig. 16) illustrates three density curves having negative, positive and zero skewnesses.

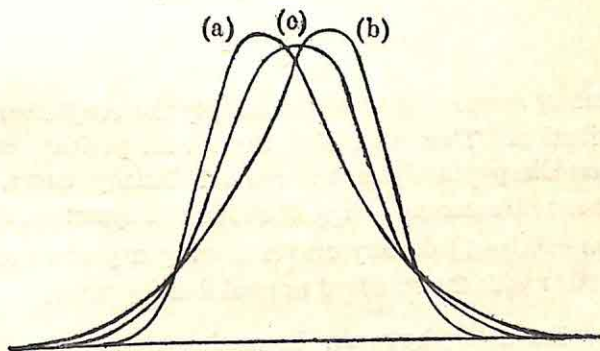


Fig. 16. (a) Negative Skewness (b) Positive Skewness (c) Symmetrical

Remark. If, for an unsymmetrical distribution, μ_3 happens to be zero, then we may take μ_5/σ^5 or, speaking generally, the first non-vanishing odd order moment of the standardised variate as a measure of skewness.

Example. GAMMA DISTRIBUTION. From Ex. 2 Sec. 7.3

$$\alpha_2 = l(l+1), \quad \alpha_3 = l(l+1)(l+2)$$

By (7.3.4)

$$\mu_3 = 2l, \quad \gamma_1 = 2l/l^{3/2} = 2/\sqrt{l}$$

7.6 FOURTH CENTRAL MOMENT

The fourth central moment μ_4 or rather the dimensionless quantity μ_4/σ^4 , which is the fourth moment of the standardised variate, is sometimes used to measure the degree of *peakedness* of the density curve of a continuous distribution near its centre (the interpretation for the discrete case in terms of the probability diagram being similar). This property is called *kurtosis*, and the *coefficient of kurtosis* β_2 is defined by

$$\beta_2 = \frac{\mu_4}{\sigma^4} \quad (7.6.1)$$

For the normal distribution $\mu_4 = 3\sigma^4$ (cf. Ex. 1 Sec. 7.3) or $\beta_2 = 3$.

Now it is customary to measure kurtosis of any distribution as compared to the normal distribution which is treated as an ideal in the field of applications, and we call the quantity

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma^4} - 3 \quad (7.6.2)$$

the *coefficient of excess of kurtosis* or simply the *coefficient of excess* of the distribution. Thus a density curve with positive excess will have a more sharp peak than the normal density curve, and the opposite for negative excess. Fig. 17 shows two symmetrical standardised (i.e. $m=0$, $\sigma=1$) density curves having negative and positive excesses together with the standard normal density curve.

Remark. It cannot, however, be mathematically proved that β_2 really gives a measure of peakedness of the density curve. On the

contrary, we can construct examples of density curves having very sharp peaks but low values of β_2 or quite flat tops with high values of β_2 . But it must be remarked that such examples are rare, and the

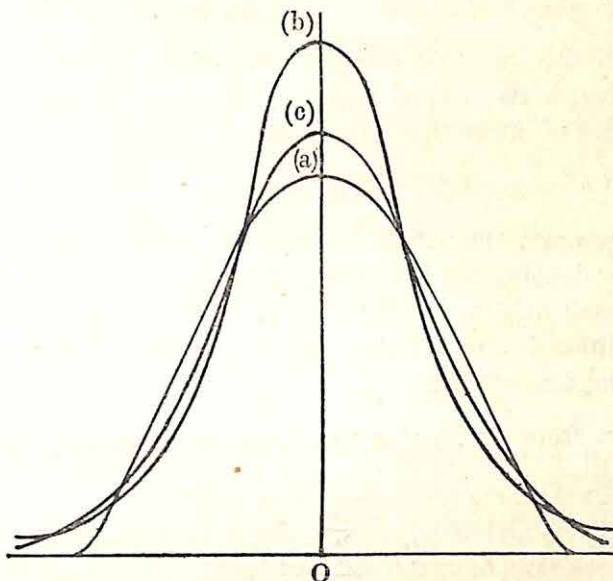


Fig. 17. (a) Negative Excess (b) Positive Excess (c) Normal Curve

above statements hold good for the majority of the distributions met in practice.

Example. GAMMA DISTRIBUTION. By (7.3.4) we have using the results of Ex. 2 Sec. 7.3

$$\begin{aligned}\mu_4 &= l(l+1)(l+2)(l+3) - 4l(l+1)(l+2)l + 6l(l+1)l^2 - 3l^4 \\ &= 3l(l+2)\end{aligned}$$

So

$$\beta_2 = 3l(l+2)/l^2 = 6/l + 3, \quad \gamma_2 = 6/l$$

7.7 MOMENT GENERATING FUNCTION

The *moment generating function* of a random variable X is a function of a real variable t denote by $\psi_x(t)$ or $\psi(t)$ and defined by

$$\psi(t) = E(e^{tX}) \quad (7.7.1)$$

Now, as to the question of convergence of the sum or integral representing $\psi(t)$, we face some difficulties. For $t=0$, $\psi(t)$ certainly exists, and $\psi(0)=1$; but for non-zero values of t the series or integral concerned does not converge for all distributions.

Differentiating the series representation of $\psi(t)$ for the discrete case term by term or its integral representation for the continuous case within the sign of integration k times at $t=0$, we get

$$\psi^{(k)}(0) = E(X^k) = a_k \quad (7.7.2)$$

assuming, however, the validity of such a process. Now a necessary condition for defining the derivatives of $\psi(t)$ at $t=0$ is that $\psi(t)$ should exist in a small neighbourhood of $t=0$. It can be proved that this simple condition is also sufficient for the existence of moments of all orders and holding of (7.7.2)

It follows from (7.7.2) that the power series development of $\psi(t)$ will be

$$\psi(t) = \sum_{k=0}^{\infty} \frac{a_k}{k!} t^k \quad (7.7.3)$$

Thus if $\psi(t)$ for any distribution is expanded by algebraic or other methods in a power series, the coefficient of t^k will be $a_k/k!$ and hence the name moment generating function.

On account of the above-mentioned convergence difficulties, the moment generating function is rather getting out of use and is conveniently replaced by what is known as the characteristic function which exists for all distributions.

7.8. CHARACTERISTIC FUNCTION

Replacing t by it ($i = \sqrt{-1}$) in $\psi(t)$, we obtain the *characteristic function* of X to be denoted by $\chi_x(t)$ or $\chi(t)$, i.e.

$$\chi(t) = E(e^{itX}) = E\{\cos(tX) + i \sin(tX)\} \quad (7.8.1)$$

The characteristic function is thus a complex-valued function of a real variable t . Since $|e^{itx}| = 1$, it may be easily seen that $\chi(t)$ exists for all values of t and for all distributions.

The existence of the characteristic function does not, however, guarantee the existence of the moments of the distribution. This we can well guess from the fact that the characteristic function exists for all distributions, but all distributions do not possess moments of all orders. But if the moment a_k exists, it can be proved that we are permitted to differentiate $E(e^{itX})$ k times at $t=0$ giving

$$\chi^{(k)}(0) = i^k a_k \quad (7.8.2)$$

Hence developing $\chi(t)$ in a power series of it , we may write formally (without any regard to its convergence)

$$\chi(t) = \sum_{k=0}^{\infty} \frac{a_k}{k!} (it)^k \quad (7.8.3)$$

so that the coefficient of $(it)^k$ in the expansion of $\chi(t)$ in powers of it is $a_k/k!$.

The most important property of the characteristic function is yet to be stated. We note that, given any distribution function $F(x)$, the corresponding characteristic function $\chi(t)$ is uniquely determined by the definition (7.8.1). Now the fundamental theorem concerning characteristic functions states that the converse is also true, i.e. the characteristic function $\chi(t)$ also uniquely determines the distribution function $F(x)$. The proof of this theorem will, however, be too difficult for us and will be omitted. But we shall make many important applications of this theorem in the following form. If, by means of some indirect method, we find the characteristic function of an unknown distribution and if this coincides with the characteristic function of a known distribution, then we can at once identify the unknown distribution with the known distribution.

The characteristic function of a continuous $g(X)$ of X may be obtained as follows. Setting $Y = g(X)$

$$\chi_y(t) = E(e^{itY}) = E\{e^{itg(X)}\} \quad (7.8.4)$$

For a linear function $aX + b$

$$\chi_{aX+b}(t) = e^{ibt} \chi_x(at) \quad (7.8.5)$$

and, in particular, for the standardised variate $X^* = (X - m)/\sigma$

$$\chi_{x^*}(t) = e^{-imt/\sigma} \chi_x(t/\sigma) \quad (7.8.6)$$

Examples

1. BINOMIAL DISTRIBUTION

$$\begin{aligned}\chi(t) &= \sum_{k=0}^n e^{itk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^n \binom{n}{k} (pe^{it})^k (1-p)^{n-k} \\ &= (pe^{it} + 1 - p)^n = (pe^{it} + q)^n \quad [\text{where } q = 1 - p]\end{aligned}$$

2. POISSON DISTRIBUTION

$$\begin{aligned}\chi(t) &= \sum_{k=0}^{\infty} e^{itk} e^{-\mu} \frac{\mu^k}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu e^{it})^k}{k!} = e^{-\mu} e^{\mu e^{it}} \\ &= e^{\mu(e^{it} - 1)}\end{aligned}$$

3. NORMAL DISTRIBUTION

$$\begin{aligned}\chi(t) &= \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} e^{itx} e^{-(x-m)^2 / 2\sigma^2} dx \\ &= e^{imt - \frac{1}{2}\sigma^2 t^2} \cdot \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} e^{-(x-m-it\sigma^2)^2 / 2\sigma^2} dx \\ &= e^{imt - \frac{1}{2}\sigma^2 t^2}\end{aligned}$$

4. GAMMA DISTRIBUTION

$$\begin{aligned}\chi(t) &= \frac{1}{\Gamma(l)} \int_0^{\infty} e^{itx} e^{-x} x^{l-1} dx = \frac{1}{\Gamma(l)} \int_0^{\infty} e^{-(1-it)x} x^{l-1} dx \\ &= \frac{1}{(1-it)^l} \cdot \frac{1}{\Gamma(l)} \int_0^{\infty} e^{-x} x^{l-1} dx \\ &= (1-it)^{-l}\end{aligned}$$

7.9 SEMI-INVARIANTS OR CUMULANTS

If $\log \chi(t)$ is developed in a power series of it in the form

$$\log \chi(t) = \sum_{k=1}^{\infty} \frac{\kappa_k}{k!} (it)^k \quad (7.9.1)$$

the coefficient κ_k is called the *semi-invariant* or *cumulant* of order k of the random variable X . We have

$$\chi(t) = 1 + \sum_{k=1}^{\infty} \frac{a_k}{k!} (it)^k$$

and hence

$$\log \chi(t) = \log \left\{ 1 + \sum_{k=1}^{\infty} \frac{a_k}{k!} (it)^k \right\}$$

Expanding this in a formal manner, we write

$$\log \chi(t) = \sum_{k=1}^{\infty} \frac{a_k}{k!} (it)^k - \frac{1}{2} \left(\sum_{k=1}^{\infty} \frac{a_k}{k!} (it)^k \right)^2 + \frac{1}{3} \left(\sum_{k=1}^{\infty} \frac{a_k}{k!} (it)^k \right)^3 - \dots$$

or

$$\begin{aligned} \log \chi(t) = & a_1(it) + (a_2 - a_1^2) \frac{(it)^2}{2!} + (a_3 - 3a_1a_2 + 2a_1^3) \frac{(it)^3}{3!} \\ & + (a_4 - 3a_2^2 - 4a_1a_3 + 12a_1^2a_2 - 6a_1^4) \frac{(it)^4}{4!} + \dots \end{aligned}$$

Comparing this expansion with (7.9.1), we obtain the following expressions of the κ_k 's in terms of the a_k 's :

$$\begin{aligned} \kappa_1 &= m \\ \kappa_2 &= a_2 - m^2 \\ \kappa_3 &= a_3 - 3a_2m + 2m^3 \\ \kappa_4 &= a_4 - 3a_2^2 - 4a_3m + 12a_2m^2 - 6m^4 \\ &\dots \dots \dots \end{aligned} \tag{7.9.2}$$

Using (7.3.4) the κ_k 's are given in terms of the μ_k 's by

$$\kappa_2 = \sigma^2, \kappa_3 = \mu_3, \kappa_4 = \mu_4 - 3\sigma^4 \text{ etc.} \tag{7.9.3}$$

Hence

$$m = \kappa_1, \sigma = \sqrt{\kappa_2}, \gamma_1 = \kappa_3/\kappa_2^{3/2}, \gamma_2 = \kappa_4/\kappa_2^2 \tag{7.9.4}$$

To find the semi-invariants of a linear function $aX + b$, we have from (7.8.5)

$$\log \chi_{aX+b}(t) = ibt + \log \chi_X(at) = ibt + \sum_{k=1}^{\infty} a_k \frac{\kappa_k}{k!} (it)^k$$

Therefore, the first semi-invariant of $aX + b$ is $a\kappa_1 + b$, and that of order $k(>1)$ is $a^k\kappa_k$.

Examples

1. BINOMIAL DISTRIBUTION

$$\begin{aligned}\log \chi(t) &= n \log (pe^{it} + q) = n \log \left\{ 1 + p it + p \frac{(it)^2}{2!} + p \frac{(it)^3}{3!} + \dots \right\} \\ &= np(it) + npq \frac{(it)^2}{2!} + npq(q-p) \frac{(it)^3}{3!} + npq(1-6pq) \frac{(it)^4}{4!} + \dots\end{aligned}$$

So

$$\kappa_1 = np, \kappa_2 = npq, \kappa_3 = npq(q-p), \kappa_4 = npq(1-6pq)$$

By (7.9.4)

$$\begin{aligned}m &= np, \quad \sigma = \sqrt{npq} \\ \gamma_1 &= \frac{q-p}{\sqrt{npq}} = \frac{1-2p}{\sqrt{npq}}, \quad \gamma_2 = \frac{1-6pq}{npq}\end{aligned}$$

2. POISSON DISTRIBUTION

$$\log \chi(t) = \mu \sum_{k=1}^{\infty} \frac{(it)^k}{k!}$$

giving $\kappa_k = \mu$ for all k . Hence

$$m = \mu, \quad \sigma = \sqrt{\mu}, \quad \gamma_1 = 1/\sqrt{\mu}, \quad \gamma_2 = 1/\mu$$

3. NORMAL DISTRIBUTION

$$\log \chi(t) = imt - \frac{1}{2}\sigma^2 t^2 = m(it) + \sigma^2 \frac{(it)^2}{2!}$$

So

$$\kappa_1 = m, \kappa_2 = \sigma^2, \kappa_3 = \kappa_4 = \dots = 0$$

which show that m is the mean, σ^2 the variance of the normal distribution and the coefficients of skewness and excess are both zero.

4. GAMMA DISTRIBUTION

$$\log \chi(t) = -l \log (1 - it) = l \sum_{k=1}^{\infty} \frac{(it)^k}{k}$$

Hence $\kappa_k = (k-1)!/l$ and

$$m = l, \quad \sigma = \sqrt{l}, \quad \gamma_1 = 2/\sqrt{l}, \quad \gamma_2 = 6/l$$

We shall now define some other useful characteristics which are not defined as mathematical expectations.

7.10 MEDIAN

This is another important measure of location denoting the point which divides the probability mass distribution into two halves. Mathematically, we define the *median* μ by the equation

$$F(\mu) = \frac{1}{2} \quad (7.10.1)$$

The median is thus the x -co-ordinate of the point of intersection of the distribution curve $y = F(x)$ with the straight line $y = \frac{1}{2}$. For a continuous distribution, since $F(x)$ is continuous and strictly monotonic increasing, a unique median exists. But for the discrete case either of the two troubles arises: (i) the straight line $y = \frac{1}{2}$ does not intersect the curve $y = F(x)$ at all, but passes between two horizontal parts of this step curve or (ii) $y = \frac{1}{2}$ intersects $y = F(x)$ at all points for which x lies in an interval $x_k \leq x < x_{k+1}$ where x_k and x_{k+1} are two consecutive points of the spectrum, so that every point of this interval satisfies equation (7.10.1) and may claim to be a median.

To cover case (i), we extend our definition of the median as follows: The median μ is a point which simultaneously satisfies the inequalities

$$F(\mu - 0) \leq \frac{1}{2}, \quad F(\mu) \geq \frac{1}{2} \quad (7.10.2)$$

The inequalities (7.10.2) have an interesting geometrical meaning, viz. the vertical steps of the distribution curve are also regarded as parts of the curve while considering its intersection with the straight line $y = \frac{1}{2}$.

To cover case (ii), we make the convention of taking the middle point of the interval, $\frac{1}{2}(x_k + x_{k+1})$ as the proper median of the distribution.

With these extensions, the median exists for every distribution and is unique.

An important property of the median is contained in the following minimum theorem: The first absolute moment about any point is minimum when taken about the median.

Proof. Let us prove the theorem for the continuous case, the proof for the discrete case being similar. If $c > \mu$

$$\begin{aligned}
 E(|X - c|) &= \int_{-\infty}^c (c - x)f(x)dx + \int_c^{\infty} (x - c)f(x)dx \\
 &= \int_{-\infty}^{\mu} (c - x)f(x)dx + \int_{\mu}^c (c - x)f(x)dx \\
 &\quad + \int_{\mu}^{\infty} (x - c)f(x)dx - \int_{\mu}^c (x - c)f(x)dx \\
 &= \int_{-\infty}^{\mu} (\mu - x)f(x)dx + \int_{\mu}^{\infty} (x - \mu)f(x)dx \\
 &\quad + (c - \mu) \left\{ \int_{-\infty}^{\mu} f(x)dx - \int_{\mu}^{\infty} f(x)dx \right\} + 2 \int_{\mu}^c (c - x)f(x)dx \\
 &= E(|X - \mu|) + (c - \mu) \{2F(\mu) - 1\} + 2 \int_{\mu}^c (c - x)f(x)dx
 \end{aligned}$$

Using (7.10.1)

$$E(|X - c|) = E(|X - \mu|) + 2 \int_{\mu}^c (c - x)f(x)dx$$

Since $c > \mu$, the integral on the R.H.S. is ≥ 0 , and hence

$$E(|X - c|) \geq E(|X - \mu|)$$

Similarly, for $c < \mu$ we have

$$E(|X - c|) = E(|X - \mu|) + 2 \int_c^{\mu} (x - c)f(x)dx$$

and the above inequality holds.

Hence $E(|X - c|) \geq E(|X - \mu|)$ always which proves the theorem.

Now $E(|X - \mu|)$ obviously gives a measure of dispersion about the median μ and is our natural choice as a dispersion characteristic when the median is selected as the characteristic of location. For a symmetrical distribution, the median lies at the point of symmetry and coincides with the mean if the latter exists.

Examples

1. Consider a discrete distribution, the spectrum of which consists of the points $0, 1, 2, \dots, n$ having the same probability mass $1/(n+1)$ at each point. If n is even, there is no point which satisfies (7.10.1), but, by the extended definition (7.10.2), the spectrum point $\frac{1}{2}n$ is the median. If n is odd, all points of the interval $\frac{1}{2}(n-1) \leq x < \frac{1}{2}(n+1)$ satisfy (7.10.1) so that, according to our convention, we take the middle point of this interval as the median, i.e. $\mu = \frac{1}{2}n$.

2. NORMAL DISTRIBUTION. It is symmetrical about the mean m , and hence $\mu = m$.

3. CAUCHY DISTRIBUTION. This distribution is also symmetrical about the point μ which shows that μ , as the notation implies, is the median. We remember that for the Cauchy distribution the mean does not exist, but the median does as it must.

7.11 MODE

Continuous case. Any point for which the density function $f(x)$ has a maximum is called a *mode* of the distribution. Now $f(x)$ may have one, two or many points of maximum, and accordingly the distribution is called *unimodal*, *bimodal* or *multimodal* respectively.

Discrete case. A point of the spectrum having the relatively tallest ordinate in the probability diagram will be called a mode, i.e. x_k is a mode if

$$f_k > f_{k-1}, f_{k+1} \quad (7.11.1)$$

Clearly, there may be more than one mode for a discrete distribution as well.

The mode is sometimes useful as a measure of location, particularly for unimodal distributions. For a unimodal symmetric distribution the mean, if it exists, and the mode are identical. Hence, for any

unimodal distribution having mode M , the quantity $m - M$ depends on the degree of asymmetry of the distribution, and we define another useful measure of skewness to be

$$\frac{m - M}{\sigma} \quad (7.11.2)$$

Examples

1. GAMMA DISTRIBUTION. Here

$$f'(x) = \frac{e^{-x} x^{l-2}}{\Gamma(l)} (l - 1 - x)$$

which vanishes for $x = l - 1$ and 0 (if $l > 2$). It may be easily seen that the maximum of $f(x)$ corresponds to $x = l - 1$. Hence the distribution is unimodal, and $M = l - 1$.

The measure of skewness (7.11.2) = $1/\sqrt{l}$, whereas the other measure $\gamma_1 = 2/\sqrt{l}$. In order to avoid such disagreements, some mathematicians take $\frac{1}{2}\gamma_1$ as a coefficient of skewness instead of γ_1 .

2. For a binomial ($2n, \frac{1}{2}$) variate $f_i = \binom{2n}{i} \left(\frac{1}{2}\right)^{2n}$ which, we know, is maximum for $i = n$, and hence $M = n$ which is also the mean of the distribution.

7.12 QUANTILES

Let p ($0 < p < 1$) be a given number. The *quantile of order p* , ξ_p will be defined by

$$F(\xi_p) = p \quad (7.12.1)$$

with extensions similar to the case of the median. Obviously $\xi_{1/2} = \mu$.

The quantiles $\xi_{1/4}$ and $\xi_{3/4}$ are much used in practice and are called the *lower* and *upper quartiles* respectively. The quantity $\frac{1}{2}(\xi_{3/4} - \xi_{1/4})$ is called the *semi-interquartile range* or the *quartile deviation* which is often used as a measure of dispersion.

The quantile of order $k/10$ is called a *decile of order k* , and the quantile of order $k/100$ a *percentile of order k* . Therefore, the 5th decile is the median, the 25th percentile the lower quartile etc. Given the deciles or better the percentiles, we can get a fairly good idea about the distribution of the probability masses.

Example. CAUCHY DISTRIBUTION. From (5.8.7)

$$F(x) = \frac{1}{\pi} \tan^{-1} \left(\frac{x - \mu}{\lambda} \right) + \frac{1}{2} = \frac{1}{4}, \frac{3}{4}$$

for $x = \mu - \lambda$, $\mu + \lambda$ respectively. Hence $\zeta_{1/4} = \mu - \lambda$, $\zeta_{3/4} = \mu + \lambda$ and the semi-interquartile range $= \lambda$.

7.13 SOME REMARKS

1. We have introduced three principal measures of location—mean, median and mode. Of these only the mean is defined as a mathematical expectation, as a result of which the rules of calculation of the mean are much simpler than those of the median or the mode. But one difficulty with the mean is that it does not always exist. Moreover, in certain distributions in which small masses occur at great distances away, the mean which denotes the centre of mass is dragged away from the bulk of the distribution. In such cases also the mean is not suitable as a measure of location.

2. If the standard deviation or the first absolute moment about the mean or the median is not available as a measure of dispersion, as, for example, in the Cauchy distribution, we may conveniently use the semi-interquartile range as a measure of dispersion. The latter also finds great use in statistical applications.

3. Although the quantities γ_1 , γ_2 and the measure of skewness (7.11.2) are dimensionless, there do not exist exact theoretical limits for these characteristics, and this is certainly a disadvantage. In practice, however, these quantities are usually found to be small.

7.14 EXERCISES

1. If n balls are drawn (a) with replacements or (b) without replacements from an urn containing N_1 white and N_2 black balls ($n \leq N_1 + N_2$), find the expectation of the number of white balls in the cases (a) and (b).

2. A point is chosen at random on a line segment AB of length $2a$. Calculate the expected values of the rectangle $AP \cdot PB$ and the difference $|AP - PB|$.

3. If X is uniformly distributed over $(0, \frac{1}{2}\pi)$, compute the expectation of the function $\sin X$. Also find the distribution of $\sin X$, and show that the mean of this distribution is the same as the above expectation.

4. In Banach's match-box problem (Ex. 7 Sec. 5.10) find the expectation of the number of matches left in one of the boxes when the other box is just found empty.

5. Show that the expectation of the number of failures preceding the first success in an infinite sequence of Bernoulli trials with probability of success p is $(1-p)/p$.

6. If X is a $\gamma(l)$ variate, compute $E(\sqrt{X})$.

7. Find the mean and variance of the rectangular distribution.

8. The *Pascal distribution* is defined by

$$x_i = i \quad (i=0, 1, 2, \dots)$$

and

$$f_i = \frac{1}{1+\mu} \left(\frac{\mu}{1+\mu} \right)^i \quad (\mu > 0)$$

Find the mean and variance of this distribution.

9. Given that the variate X is normal $(0, 1)$, find the variance of e^X .

10. The first, second and third moments of a probability distribution about the point 2 are 1, 16, -40 respectively. Find the mean, variance and the third central moment.

11. The probability density of a continuous distribution is given by : $f(x) = \frac{3}{2}x(2-x)$ ($0 < x < 2$). Compute the mean, variance and the coefficient of skewness γ_1 .

12. For the binomial (n, p) distribution, prove that

$$\mu_{k+1} = p(1-p) \left(nk\mu_{k-1} + \frac{d\mu_k}{dp} \right)$$

and hence obtain γ_1 and γ_2 .

13. Prove that $E(X^2) \geq \{E(X)\}^2$. Deduce that the first absolute moment about the mean is at most equal to the standard deviation.

14. For the Poisson distribution with parameter μ , prove that

$$\mu_{k+1} = \mu \left(k\mu_{k-1} + \frac{d\mu_k}{d\mu} \right)$$

Hence calculate γ_1 and γ_2 .

15. Show that the first absolute moment about the mean for the normal (m, σ^2) distribution is $\sqrt{2/\pi} \sigma$.

16. Calculate the k th moment (about the origin) for a $\beta_1(l, m)$ distribution, and hence obtain the variance. Show also that, for $l, m > 1$, there exists a unique mode having the value $(l-1)/(l+m-2)$.

17. A continuous distribution is given by

$$f(x) = \frac{1}{x\sqrt{2\pi}} e^{-(\log x)^2/2} \quad (x > 0)$$

$$= 0 \quad (x < 0)$$

which is called a *log-normal distribution*. For this distribution calculate the mean, mode and standard deviation, and obtain the coefficient of skewness (7.11.2)

18. Show that the mode M of the Poisson distribution with mean μ is the integer(s) determined by the inequalities: $\mu - 1 \leq M \leq \mu$.

19. A continuous distribution has probability density $f(x) = ae^{-ax}$ ($0 < x < \infty$; $a > 0$). Calculate the moment generating function, and hence obtain a_k .

20. Prove that the moment generating function of a uniform distribution over the interval $(-a, a)$ is $\sinh at/at$. Hence calculate the central moments.

21. Show that the characteristic function for the Pascal distribution defined in Ex. 8 is $[1 - \mu(e^{it} - 1)]^{-1}$. By expanding the characteristic function in powers of it find a_1, a_2, a_3, a_4 , and hence calculate the coefficient of skewness γ_1 and the coefficient of excess γ_2 .

Also calculate the first four cumulants of the distribution and verify therefrom the values of γ_1 and γ_2 .

22. Find the first four cumulants of the *Laplace distribution* defined by

$$f(x) = \frac{1}{2\lambda} e^{-|x - \mu|/\lambda} \quad (-\infty < x < \infty; \lambda > 0)$$

and hence find the values of m, σ, γ_1 and γ_2 .

23. Find the mean, median and the mode of a binomial $(4, \frac{1}{2})$ variate.
24. Find the median for the Poisson distribution having mean 2.
25. Find the lower and upper quartiles for the distribution of the number of points on a card drawn at random from a full pack, and calculate the semi-inter-quartile range. (Take 11 points for the jack, 12 for the queen and 13 for the king.)
26. Calculate the first absolute moment about the mean and the semi-inter-quartile range for the Laplace distribution defined in Ex. 22.

MATHEMATICAL EXPECTATIONS II

A. TWO-DIMENSIONAL CASE

8.1 EXPECTATION FOR A BIVARIATE DISTRIBUTION

Consider the joint distribution of two random variables X and Y . The *expectation* or the *mean value* of a continuous function $g(X, Y)$ of X, Y is defined by

$$E\{g(X, Y)\} = \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} g(x_i, y_j) f_{ij} \quad \text{for the discrete case} \quad (8.1.1)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \quad \text{for the continuous case}$$

provided the series or integral is absolutely convergent.

Remark. Now the mean value $E\{g(X)\}$ may be calculated in two ways, viz. (i) by formula (7.1.1) with respect to the distribution of X alone and (ii) by the above formula (8.1.1) with respect to the joint distribution of X with any other random variable Y , and it needs proving that these two values are the same so that our definitions are consistent.

Proof. DISCRETE CASE. According to (8.1.1)

$$\begin{aligned} E\{g(X)\} &= \sum_i \sum_j g(x_i) f_{ij} = \sum_i g(x_i) \sum_j f_{ij} \\ &= \sum_i g(x_i) f_{i.} = \sum_i g(x_i) f_{xi} \end{aligned}$$

which is the value given by (7.1.1).

CONTINUOUS CASE. Here

$$\begin{aligned} E\{g(X)\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} g(x) dx \int_{-\infty}^{\infty} f(x, y) dy = \int_{-\infty}^{\infty} g(x) f_x(x) dx \end{aligned}$$

It follows from the above remark that the characteristics such as the mean, variance etc. of the random variables X and Y are uniquely defined whether they are calculated from their individual distributions or from their joint distribution.

Geometrically, the point (m_x, m_y) represents the centre of mass of the two-dimensional probability mass distribution; the variances σ_x^2 , σ_y^2 represent respectively the moments of inertia of the mass distribution about the lines $x = m_x$ and $y = m_y$ which are parallel to the axes passing through the centre of mass and are thus measures of dispersion about the said lines.

An obvious but important property of expectations is

$$\begin{aligned} E\{g_1(X, Y) + g_2(X, Y) + \dots + g_n(X, Y)\} \\ = E\{g_1(X, Y)\} + E\{g_2(X, Y)\} + \dots + E\{g_n(X, Y)\} \end{aligned} \quad (8.1.2)$$

provided all the expectations on the R.H.S exist (which implies the existence of the L.H.S.).

In particular, we have

$$E(X + Y) = E(X) + E(Y) \quad (8.1.3)$$

(8.1.3) gives the *addition rule for mean values* which states that if the mean values of X and Y exist, then the mean value of their sum $X + Y$ also exists and is equal to the sum of their mean values.

Examples

1. If X, Y are random variables defined in Ex. 1 Sec. 6.2, compute $E(|X - Y|)$.

Here $(x_i, y_j) = (i, j)$ ($i = 0, 1, 2$; $j = 0, 1, 2, 3$) and $f_{ij} = 1/9$ for all i, j except that $f_{13} = f_{22} = f_{23} = 0$. Hence

$$\begin{aligned} E(|X - Y|) &= \sum_{j=0}^3 \sum_{i=0}^2 |i - j| f_{ij} \\ &= \frac{1}{9} (|0 - 0| + |0 - 1| + |0 - 2| + |0 - 3| + |1 - 0| + \\ &\quad + |1 - 1| + |1 - 2| + |2 - 0| + |2 - 1|) \\ &= 11/9 \end{aligned}$$

2. If X, Y are independent standard normal variates, find the mean value of the greater of $|X|$ and $|Y|$.

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2 + y^2)/2} \quad (-\infty < x < \infty, -\infty < y < \infty)$$

$$E\{\max(|X|, |Y|)\}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max(|x|, |y|) f(x, y) dx dy$$

$$= \frac{1}{2\pi} \int \int_{|x| > |y|} |x| e^{-(x^2 + y^2)/2} dx dy + \frac{1}{2\pi} \int \int_{|y| > |x|} |y| e^{-(x^2 + y^2)/2} dx dy$$

$$= \frac{1}{\pi} \int \int_{|x| > |y|} |x| e^{-(x^2 + y^2)/2} dx dy \quad [\text{from symmetry}]$$

$$= \frac{2}{\pi} \int_{-\infty}^{\infty} e^{-y^2/2} dy \int_{|y|}^{\infty} x e^{-x^2/2} dx \quad [\text{from symmetry}]$$

$$= \frac{2}{\pi} \int_{-\infty}^{\infty} e^{-y^2} dy = \frac{2}{\sqrt{\pi}}$$

8.2 MOMENTS

We define the *moments* (about the origin) of the joint distribution of X and Y by

$$a_{kl} = E(X^k Y^l) \quad (8.2.1)$$

where k, l are non-negative integers; a_{kl} is called a moment of order $k+l$. We have

$$a_{k0} = a_{xk}, \quad a_{0l} = a_{yl}$$

In particular

$$a_{00} = 1, \quad a_{10} = m_x, \quad a_{01} = m_y$$

The central moments are given by

$$\mu_{kl} = E\{(X - m_x)^k (Y - m_y)^l\} \quad (8.2.2)$$

Hence

$$\mu_{k0} = \mu_{xk}, \quad \mu_{0l} = \mu_{yl}$$

$$\mu_{00} = 1, \quad \mu_{10} = 0, \quad \mu_{01} = 0, \quad \mu_{20} = \sigma_x^2, \quad \mu_{02} = \sigma_y^2$$

8.3 COVARIANCE, CORRELATION COEFFICIENT

The second order mixed central moment μ_{11} furnishes an important measure of, what we may roughly say, the jointness of the bivariate distribution and is called the *covariance of X and Y* , to be denoted by $\text{cov}(X, Y)$, i.e.

$$\text{cov}(X, Y) = \mu_{11} = E\{(X - m_x)(Y - m_y)\} \quad (8.3.1)$$

Now if we want to find a dimensionless measure of the property expressed by the covariance, we have to introduce, following our usual practice, the standardised random variables X^*, Y^* in places of X, Y respectively and take $\text{cov}(X^*, Y^*)$ as the required measure. We shall call $\text{cov}(X^*, Y^*)$ the *correlation coefficient* of X and Y and denote it by $\rho(X, Y)$ or ρ_{xy} or simply ρ , i.e.

$$\begin{aligned} \rho(X, Y) &= \text{cov}(X^*, Y^*) = E(X^* Y^*) \\ &= \frac{E\{(X - m_x)(Y - m_y)\}}{\sigma_x \sigma_y} \\ &= \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \quad (\sigma_x, \sigma_y \neq 0) \end{aligned} \quad (8.3.2)$$

1. If $a_1 (\neq 0)$, $a_2 (\neq 0)$, b_1 , b_2 are constants, then

$$\rho(a_1 X + b_1, a_2 Y + b_2) = \frac{a_1 a_2}{|a_1| |a_2|} \rho(X, Y) \quad (8.3.3)$$

Proof. $(a_1 X + b_1)^* = \frac{a_1}{|a_1|} X^*$, $(a_2 Y + b_2)^* = \frac{a_2}{|a_2|} Y^*$

So

$$\begin{aligned} \rho(a_1 X + b_1, a_2 Y + b_2) &= E \left(\frac{a_1 a_2}{|a_1| |a_2|} X^* Y^* \right) \\ &= \frac{a_1 a_2}{|a_1| |a_2|} E(X^* Y^*) = \text{R.H.S. of (8.3.3)} \end{aligned}$$

Hence if $a_1, a_2 > 0$

$$\rho(a_1 X + b_1, a_2 Y + b_2) = \rho(X, Y) \quad (8.3.4)$$

This shows that the correlation coefficient is independent of the choice of origins and units of measurements of the random variables. In particular, $\rho(X^*, Y^*) = \rho(X, Y)$, since $\sigma_x, \sigma_y > 0$.

2. $-1 \leq \rho(X, Y) \leq 1 \quad (8.3.5)$

Proof. $0 \leq (X^* \pm Y^*)^2 = X^{*2} + Y^{*2} \pm 2X^* Y^*$

Considering expectations and remembering that $E(X^{*2}) = E(Y^{*2}) = 1$, we get

$$0 \leq E\{(X^* \pm Y^*)^2\} = 2\{1 \pm \rho(X, Y)\}$$

which gives (8.3.5).

If $\rho(X, Y) = \pm 1$, we must have $X^* \mp Y^* = 0$ or $Y^* = \pm X^*$ or

$$\frac{Y - m_y}{\sigma_y} = \pm \frac{X - m_x}{\sigma_x} \quad (8.3.6)$$

Thus if $\rho(X, Y) = \pm 1$, Y is a linear function of X given by (8.3.6), or in other words, the whole probability mass of the bivariate distribution is situated on the straight line

$$\frac{y - m_y}{\sigma_y} = \pm \frac{x - m_x}{\sigma_x}$$

Conversely, if $Y = aX + b$, a, b being constants, $\rho(X, Y) = \pm 1$, for

$$\rho(X, aX + b) = \frac{a}{|a|} \rho(X, X) = \pm 1$$

3. If $\rho(X, Y) = 0$, we cannot, however, conclude that X and Y are independent. In that case we shall simply say that X and Y are *uncorrelated*. The detailed discussion regarding this will be taken up in Sec. 8.7.

4. Let us calculate the variance of $X \pm Y$. We have $E(X \pm Y) = m_x \pm m_y$, and

$$\begin{aligned}\{(X \pm Y - (m_x \pm m_y))\}^2 &= \{(X - m_x) \pm (Y - m_y)\}^2 \\ &= (X - m_x)^2 + (Y - m_y)^2 \pm 2(X - m_x)(Y - m_y)\end{aligned}$$

Taking expectations of both sides, we get the *variance formula* :

$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2 \text{cov}(X, Y) \quad (8.3.7)$$

which, in another form, is

$$\sigma^2(X \pm Y) = \sigma^2(X) + \sigma^2(Y) \pm 2\sigma(X)\sigma(Y)\rho(X, Y) \quad (8.3.8)$$

If X and Y are uncorrelated

$$\sigma^2(X \pm Y) = \sigma^2(X) + \sigma^2(Y) \quad (8.3.9)$$

$$5. \quad \mu_{11} = a_{11} - m_x m_y \quad (8.3.10)$$

Proof. $(X - m_x)(Y - m_y) = XY - m_y X - m_x Y + m_x m_y$. Hence

$$\mu_{11} = E(XY) - m_y E(X) - m_x E(Y) + m_x m_y = a_{11} - m_x m_y$$

We shall now prove an interesting theorem.

Theorem. If, for any pair of correlated random variables X and Y , we make a linear transformation $(X, Y) \rightarrow (U, V)$ given by a rotation of the axes through a constant angle α , i.e.

$$U = X \cos \alpha + Y \sin \alpha, \quad V = -X \sin \alpha + Y \cos \alpha \quad (8.3.11)$$

then U and V will be uncorrelated if α is given by

$$\tan 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \quad (8.3.12)$$

where $\rho = \rho(X, Y)$.

Proof. $m_u = m_x \cos \alpha + m_y \sin \alpha$, $m_v = -m_x \sin \alpha + m_y \cos \alpha$

So

$$\begin{aligned}(U - m_u)(V - m_v) &= \{(X - m_x) \cos \alpha + (Y - m_y) \sin \alpha\} \\ &\quad \times \{- (X - m_x) \sin \alpha + (Y - m_y) \cos \alpha\} \\ &= -\frac{1}{2}\{(X - m_x)^2 - (Y - m_y)^2\} \sin 2\alpha + (X - m_x)(Y - m_y) \cos 2\alpha\end{aligned}$$

Hence

$$\text{cov}(U, V) = -\frac{1}{2}(\sigma_x^2 - \sigma_y^2) \sin 2\alpha + \rho \sigma_x \sigma_y \cos 2\alpha$$

which vanishes if (8.3.12) holds. Hence the theorem.

Examples

1. In Ex. 1 Sec. 6.2 calculate m_x , m_y , σ_x^2 , σ_y^2 and ρ .

$$m_x = \sum_{j=0}^3 \sum_{i=0}^2 i f_{ij} = \frac{7}{9}, \quad a_{x2} = \sum_{j=0}^3 \sum_{i=0}^2 i^2 f_{ij} = \frac{11}{9}$$

so that $\sigma_x^2 = \frac{50}{81}$. These may also be calculated from the marginal distribution of X . Similarly

$$m_y = \frac{10}{9}, \quad a_{y2} = \frac{20}{9}, \quad \sigma_y^2 = \frac{80}{81}, \quad a_{11} = \sum_{j=0}^3 \sum_{i=0}^2 i j f_{ij} = \frac{5}{9}$$

Hence by (8.3.10)

$$\mu_{11} = -25/81, \quad \rho = -\sqrt{10/8}$$

2. BIVARIATE NORMAL DISTRIBUTION. Since the (marginal) distribution of X and Y are normal (m_x, σ_x) and (m_y, σ_y) respectively, the parameters m_x , m_y , σ_x , σ_y have their natural significances, and we have

$$\begin{aligned} \text{cov}(X, Y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \\ &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-m_x)(y-m_y) e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right\}} dx dy \\ &= \frac{\sigma_x\sigma_y}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy e^{-(x^2-2\rho xy+y^2)/2(1-\rho^2)} dx dy \end{aligned}$$

or

$$\begin{aligned} \rho(X, Y) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2} dy \left\{ \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} x e^{-(x-\rho y)^2/2(1-\rho^2)} dx \right\} \\ &= \rho \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy = \rho \end{aligned}$$

Thus the parameter ρ denotes the correlation coefficient of X and Y .

8.4 CHARACTERISTIC FUNCTION

The characteristic function $\chi(t, u)$ of the joint distribution of X, Y is defined by

$$\chi(t, u) = E\{e^{i(tX + uY)}\} \quad (8.4.1)$$

We note $\chi(t, 0) = \chi_x(t)$, $\chi(0, u) = \chi_y(u)$, $\frac{\partial^2 \chi}{\partial t \partial u} \Big|_{(0,0)} = i^2 a_{11}$ etc. Therefore, the development of $\chi(t, u)$ in powers of it and iu will be

$$\begin{aligned} \chi(t, u) = 1 + (a_{x1}it + a_{y1}iu) + \frac{1}{2!} \{a_{x2}(it)^2 + 2a_{11}(it)(iu) \\ + a_{y2}(iu)^2\} + \dots \end{aligned} \quad (8.4.2)$$

8.5 SOME EXTENSIONS TO n -DIMENSIONS

For any set of n random variables X_1, X_2, \dots, X_n having means m_1, m_2, \dots, m_n and standard deviations $\sigma_1, \sigma_2, \dots, \sigma_n$ respectively, the mean M_n and standard deviation Σ_n of their sum

$$S_n = X_1 + X_2 + \dots + X_n \quad (8.5.1)$$

are obtained by formal generalisations of (8.1.3) and (8.3.8) which give

$$M_n = m_1 + m_2 + \dots + m_n \quad (8.5.2)$$

$$\Sigma_n^2 = \sum_{i=1}^n \sigma_i^2 + 2 \sum_{i < j} \sigma_i \sigma_j \rho(X_i, X_j) \quad (8.5.3)$$

(i, j being any combination of $1, 2, \dots, n$ taken 2 at a time in the second summation.)

If X_1, X_2, \dots, X_n are *pairwise uncorrelated*, we get the simple variance formula

$$\Sigma_n^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2 \quad (8.5.4)$$

More generally, for a linear combination

$$X = a_1 X_1 + a_2 X_2 + \dots + a_n X_n \quad (8.5.5)$$

we have using (7.4.5)

$$m_x = a_1 m_1 + a_2 m_2 + \dots + a_n m_n \quad (8.5.6)$$

$$\sigma_x^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 + 2 \sum_{i < j} a_i a_j \sigma_i \sigma_j \rho(X_i, X_j) \quad (8.5.7)$$

If $\rho(X_i, X_j) = 0$ ($i \neq j$), the last equation reduces to

$$\sigma_x^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2 \quad (8.5.8)$$

Example. DRAWINGS WITHOUT REPLACEMENT. If n balls are drawn successively without replacements from an urn containing N_1 white and N_2 black balls ($n \leq N_1 + N_2$), find the mean and variance of the number of white balls.

Instead of calculating the mean and variance directly from the distribution of the number of white balls, let us here follow an indirect method which will illustrate the use of the above formulæ. We define n random variables X_1, X_2, \dots, X_n on the event space of n drawings by; $X_i = 0$ or 1 corresponding to black or white ball in the i th trial so that S_n denotes the number of white balls.

From the symmetry of the situation, while considering the i th drawing we may forget about the rest of the drawings, and as such

$$P(X_i = 1) = N_1 / (N_1 + N_2)$$

So

$$m_i = E(X_i) = N_1 / (N_1 + N_2), \quad E(X_i^2) = N_1 / (N_1 + N_2)$$

Hence by (7.4.3)

$$\sigma_i^2 = \frac{N_1 N_2}{(N_1 + N_2)^2}$$

By (8.5.2)

$$M_n = \frac{n N_1}{N_1 + N_2}$$

Now consider the distribution of the two-dimensional variate (X_i, X_j) ($i \neq j$); its spectrum consists of the four points $(0, 0)$, $(1, 0)$, $(0, 1)$, $(1, 1)$, the probability mass at the last point being

$$\frac{N_1(N_1 - 1)}{(N_1 + N_2)(N_1 + N_2 - 1)}. \quad \text{Then}$$

$$E(X_i X_j) = \frac{N_1(N_1 - 1)}{(N_1 + N_2)(N_1 + N_2 - 1)}$$

and by (8.3.10)

$$\text{cov}(X_i, X_j) = -\frac{N_1 N_2}{(N_1 + N_2)^2 (N_1 + N_2 - 1)}$$

From (8.5.3)

$$\Sigma_n^2 = \frac{n N_1 N_2}{(N_1 + N_2)^2} - 2 \binom{n}{2} \frac{N_1 N_2}{(N_1 + N_2)^2 (N_1 + N_2 - 1)}$$

or

$$\Sigma_n^2 = \frac{n N_1 N_2}{(N_1 + N_2)^2} \left(1 - \frac{n-1}{N_1 + N_2 - 1} \right) \quad (8.5.9)$$

Remark. In the case of drawings with replacement, the number of white balls has, we know, a binomial distribution with parameters $\{n, N_1/(N_1 + N_2)\}$. Hence

$$M_n = \frac{n N_1}{N_1 + N_2}, \quad \Sigma_n^2 = \frac{n N_1 N_2}{(N_1 + N_2)^2}$$

so that the mean remains the same but the variance is different.

The joint characteristic function of X_1, X_2, \dots, X_n will be given by

$$\chi(t_1, t_2, \dots, t_n) = E\{e^{i(t_1 X_1 + t_2 X_2 + \dots + t_n X_n)}\} \quad (8.5.10)$$

B. INDEPENDENT RANDOM VARIABLES

8.6. MULTIPLICATION RULE FOR EXPECTATIONS

Theorem. If X and Y are independent random variables and $g_1(X)$ and $g_2(Y)$ are continuous functions of X and Y respectively whose expectations exist, then

$$E\{g_1(X)g_2(Y)\} = E\{g_1(X)\} E\{g_2(Y)\} \quad (8.6.1)$$

Proof. DISCRETE CASE

$$E\{g_1(X)\} = \sum_i g_1(x_i) f_{xi}, \quad E\{g_2(Y)\} = \sum_j g_2(y_j) f_{yj}$$

If $E\{g_1(X)\}$ and $E\{g_2(Y)\}$ exist, the series representing these are absolutely convergent, and therefore

$$\left\{ \sum_i g_1(x_i) f_{xi} \right\} \left\{ \sum_j g_2(y_j) f_{yj} \right\} = \sum_i \sum_j g_1(x_i) g_2(y_j) f_{xi} f_{yj}$$

and the series on the R.H.S. is also absolutely convergent.

Since X and Y are independent, $f_{xi}f_{yj}=f_{ij}$ for all i, j . So

$$\begin{aligned} E\{g_1(X)\} E\{g_2(Y)\} &= \sum_i \sum_j g_1(x_i)g_2(y_j)f_{xi}f_{yj} \\ &= \sum_i \sum_j g_1(x_i)g_2(y_j)f_{ij} \\ &= E\{g_1(X)g_2(Y)\} \end{aligned}$$

This shows that $E\{g_1(X)g_2(Y)\}$ exists, and formula (8.6.1) holds.

CONTINUOUS CASE. The proof is similar to the discrete case. We have

$$E\{g_1(X)\} = \int_{-\infty}^{\infty} g_1(x)f_x(x)dx, \quad E\{g_2(Y)\} = \int_{-\infty}^{\infty} g_2(y)f_y(y)dy$$

the integrals being absolutely convergent. Hence

$$\left\{ \int_{-\infty}^{\infty} g_1(x)f_x(x)dx \right\} \left\{ \int_{-\infty}^{\infty} g_2(y)f_y(y)dy \right\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(x)g_2(y)f_x(x)f_y(y)dx dy$$

and the double integral is absolutely convergent.

For independence of X and Y , we have $f_x(x)f_y(y)=f(x, y)$. Inserting this in the above equation, we get the theorem.

A particular form of (8.6.1) is

$$E(XY) = E(X) E(Y) \quad (8.6.2)$$

which is called the *multiplication rule for mean values*. Stated completely, the multiplication rule says that if X and Y are independent random variables having existent means, then the mean of their product XY exists and is equal to the product of their means.

This theorem may be easily generalised to more than two variates. For three variates X, Y, Z , if (X, Y) and Z are independent, we shall have

$$E\{g_1(X, Y)g_2(Z)\} = E\{g_1(X, Y)\}E\{g_2(Z)\} \quad (8.6.3)$$

provided the expectations on the R. H. S. exist.

If X, Y, Z are mutually independent, it follows from Theorem II Sec. 6.7 that (X, Y) and Z are independent so that (8.6.3) holds, and further

$$E\{g_1(X)g_2(Y)g_3(Z)\} = E\{g_1(X)\} E\{g_2(Y)\} E\{g_3(Z)\} \quad (8.6.4)$$

Generalising to the n -variate case we have the following result : If X_1, X_2, \dots, X_n are mutually independent, then by Theorem III (b) Sec. 6.7 the random variables

$$(X_1, \dots, X_{k_1}), (X_{k_1+1}, \dots, X_{k_2}), \dots, (X_{k_{m+1}+1}, \dots, X_n)$$

where $1 < k_1 < k_2 < \dots < k_m < n$ are also mutually independent, and hence

$$\begin{aligned} E\{g_1(X_1, \dots, X_{k_1})g_2(X_{k_1+1}, \dots, X_{k_2}) \dots g_{m+1}(X_{k_{m+1}+1}, \dots, X_n)\} \\ = E\{g_1(X_1, \dots, X_{k_1})\} E\{g_2(X_{k_1+1}, \dots, X_{k_2})\} \dots E\{g_{m+1}(X_{k_{m+1}+1}, \dots, X_n)\} \end{aligned} \quad (8.6.5)$$

where g 's denote continuous functions of their arguments, provided all the expectations on the R.H.S. exist.

In particular, we have the simple formula

$$E\{g_1(X_1)g_2(X_2) \dots g_n(X_n)\} = E\{g_1(X_1)\} E\{g_2(X_2)\} \dots E\{g_n(X_n)\} \quad (8.6.6)$$

8.7 MOMENTS

If X, Y are independent, it follows from (8.6.1) that

$$a_{kl} = E(X^k) E(Y^l) = a_{xk} a_{yl} \quad (8.7.1)$$

Similarly

$$\mu_{kl} = \mu_{xk} \mu_{yl} \quad (8.7.2)$$

Hence $\mu_{11} = \mu_{x1} \mu_{y1} = 0$ or $\rho(X, Y) = 0$.

Thus if X, Y are independent, they are necessarily uncorrelated. But the converse of this is not true; the random variables may be dependent, but their correlation coefficient may vanish due to symmetry of the distribution. The following example is to the point.

Example. Let X have any distribution symmetrical about the origin. Then $m_x = E(X) = 0$, $E(X^3) = 0$. Setting $Y = X^2$

$$\text{cov}(X, Y) = E\{X(Y - m_y)\} = E\{X(X^2 - m_y)\} = 0$$

i.e. X, Y are uncorrelated, although X, Y are even functionally dependent.

Remark. If X_1, X_2, \dots, X_n are mutually independent, they are certainly pairwise independent (cf. Theorem III (a) Sec. 6.7) so that they are pairwise uncorrelated, and hence the simplified variance formulæ (8.5.4) and (8.5.8) hold for these variates.

8.8 CHARACTERISTIC FUNCTION

For independent X, Y

$$\chi(t, u) = E(e^{itX} e^{iuY}) = E(e^{itX}) E(e^{iuY})$$

or

$$\chi(t, u) = \chi_x(t) \chi_y(u) \quad (8.8.1)$$

Conversely, if (8.8.1) holds, then X, Y can be proved to be independent. The proof of the converse is, however, beyond the level of this book. Taking this for granted, we have, generalising for n variates, the following important theorem.

Theorem. A necessary and sufficient condition for the random variables X_1, X_2, \dots, X_n to be mutually independent is that their joint characteristic function is given by

$$\chi(t_1, t_2, \dots, t_n) = \chi_1(t_1) \chi_2(t_2) \dots \chi_n(t_n) \quad (8.8.2)$$

where $\chi_1(t_1), \chi_2(t_2), \dots, \chi_n(t_n)$ respectively denote the characteristic functions of X_1, X_2, \dots, X_n .

We may now suggest a simple proof of Theorem III(c) Sec. 6.7.

Setting $Y_1 = g_1(X_1, \dots, X_{k_1}), Y_2 = g_2(X_{k_1+1}, \dots, X_{k_2}), Y_{m+1} = g_{m+1}(X_{k_{m+1}+1}, \dots, X_n)$, we have

$$\begin{aligned} & \chi_{y_1, y_2, \dots, y_{m+1}}(t_1, t_2, \dots, t_{m+1}) \\ &= E\{e^{it_1 Y_1 + t_2 Y_2 + \dots + t_{m+1} Y_{m+1}}\} \\ &= E\{e^{it_1 g_1} e^{it_2 g_2} \dots e^{it_{m+1} g_{m+1}}\} \\ &= E(e^{it_1 g_1}) E(e^{it_2 g_2}) \dots E(e^{it_{m+1} g_{m+1}}) \quad [\text{by (8.6.5)}] \\ &= E(e^{it_1 Y_1}) E(e^{it_2 Y_2}) \dots E(e^{it_{m+1} Y_{m+1}}) \\ &= \chi_{y_1}(t_1) \chi_{y_2}(t_2) \dots \chi_{y_{m+1}}(t_{m+1}) \end{aligned}$$

Hence, by the above theorem, Y_1, Y_2, \dots, Y_{m+1} are mutually independent.

Sum of independent random variables. Let X_1, X_2, \dots, X_n be mutually independent random variables having characteristic functions $\chi_1(t), \chi_2(t), \dots, \chi_n(t)$ respectively. Then the characteristic function $K(t)$ of their sum S_n is given by

$$\begin{aligned} K(t) &= E(e^{itS_n}) = E\{e^{it(X_1 + X_2 + \dots + X_n)}\} \\ &= E(e^{itX_1} e^{itX_2} \dots e^{itX_n}) \\ &= E(e^{itX_1}) E(e^{itX_2}) \dots E(e^{itX_n}) \end{aligned} \quad [\text{by (8.6.5)}]$$

or

$$K(t) = \chi_1(t) \chi_2(t) \dots \chi_n(t) \quad (8.8.3)$$

(8.8.3) states an important property of characteristic functions, viz. the characteristic function of a sum of mutually independent random variables is the product of their individual characteristic functions.

For a linear combination $X = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$, we get using (7.8.5)

$$\chi_x(t) = \chi_1(a_1 t) \chi_2(a_2 t) \dots \chi_n(a_n t) \quad (8.8.4)$$

REPRODUCTIVE PROPERTIES OF VARIOUS DISTRIBUTIONS

The *reproductive properties* of different distributions may be very easily established by making use of the characteristic functions together with the fact that a characteristic function uniquely determines the distribution.

1. If X_1, X_2, \dots, X_n are mutually independent binomial variates having parameters $(v_1, p), (v_2, p), \dots, (v_n, p)$ respectively, then their sum S_n is also binomially distributed with parameters (v, p) , where $v = v_1 + v_2 + \dots + v_n$.

Proof. The characteristic function of X_k , $\chi_k(t) = (pe^{it} + q)^{v_k}$ ($k = 1, 2, \dots, n$). Hence by (8.8.3), $K(t) = (pe^{it} + q)^v$ which is indeed the characteristic function of the binomial (v, p) distribution. Hence we conclude that S_n is binomial (v, p) .

2. If X_1, X_2, \dots, X_n are mutually independent Poisson variates having parameters $\mu_1, \mu_2, \dots, \mu_n$ respectively, then their sum S_n is a Poisson- $(\mu_1 + \mu_2 + \dots + \mu_n)$ variate.

Proof. Here $\chi_k(t) = e^{\mu_k(e^{it} - 1)}$ ($k = 1, 2, \dots, n$) so that

$$K(t) = e^{(\sum \mu_k)(e^{it} - 1)}$$

Hence the result.

3. If X_1, X_2, \dots, X_n are mutually independent normal variates having means m_1, m_2, \dots, m_n and standard deviations $\sigma_1, \sigma_2, \dots, \sigma_n$ respectively, then any linear combination $X = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$ is also normally distributed whose mean and standard deviation are given by

$$m_x = a_1 m_1 + a_2 m_2 + \dots + a_n m_n$$

$$\sigma_x^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2$$

Proof. $\chi_k(t) = e^{im_k t - \frac{1}{2}\sigma_k^2 t^2}$, and so

$$\chi_k(a_k t) = e^{ia_k m_k t - \frac{1}{2}a_k^2 \sigma_k^2 t^2} \quad (k = 1, 2, \dots, n)$$

By (8.8.4)

$$\chi_x(t) = e^{im_x t - \frac{1}{2}\sigma_x^2 t^2}$$

which proves the theorem.

Remark. The above result is slightly more general than the reproductive theorem for the normal distribution; the latter is obtained from the former by putting $a_1 = a_2 = \dots = a_n = 1$.

4. If X_1, X_2, \dots, X_n are mutually independent gamma variates having parameters l_1, l_2, \dots, l_n respectively, their sum is a $\gamma(l_1 + l_2 + \dots + l_n)$ variate.

Proof. $\chi_k(t) = (1 - it)^{-l_k}$ ($k = 1, 2, \dots, n$). Rest is obvious.

8.9 ANOTHER DISCUSSION ON BERNOULLI TRIALS

As in the example of Sec. 8.5, we define n random variables X_1, X_2, \dots, X_n on the event space S_n of a sequence of n Bernoulli trials as follows: X_i takes the value—0 or 1 corresponding to event points for which the i th trial results in a failure or success respectively ($i = 1, 2, \dots, n$). Hence

$$P(X_i = 0) = q, \quad P(X_i = 1) = p$$

i.e. X_i is binomial (1, p).

The spectrum of the n -dimensional variate (X_1, X_2, \dots, X_n) then consists of the 2^n points

$$(i_1, i_2, \dots, i_n) \quad (i_1, i_2, \dots, i_n = 0, 1)$$

each of which corresponds to an event point of S_n . The definition of independence of the n trials gives (cf. Sec. 4.3)

$P(X_1 = i_1, X_2 = i_2, \dots, X_n = i_n) = P(X_1 = i_1) P(X_2 = i_2) \dots P(X_n = i_n)$
which shows that X_1, X_2, \dots, X_n are mutually independent.

Now since X_1, X_2, \dots, X_n are mutually independent variates, each binomial $(1, p)$, it follows from the reproductive property of the binomial distribution that their sum $S_n = X_1 + X_2 + \dots + X_n$, which represents the number of successes in n trials, has a binomial (n, p) distribution.

Example. RANDOM WALK PROBLEM. This is an interesting model connected with a Bernoullian sequence of trials. Let a particle be initially at a point r , a given positive integer, on the x -axis. Now a sequence of n Bernoulli trials are performed, and for each trial the particle moves by jump through unit distance either in the forward or the backward direction, according as the result of the trial is a success or a failure. Our problem will be to find the probability distribution of the co-ordinate of the particle after n trials or jumps.

Let the random variable X_i' denote the displacement of the particle at the i th jump ($i = 1, 2, \dots, n$). Then X_i' takes the two values -1 and 1 with probabilities q and p respectively. We note that X_i' does not exactly have the two-point binomial distribution, but the variate $X_i = (X_i' + 1)/2$ whose spectrum consists of the points 0 and 1 is indeed binomial $(1, p)$. After n jumps the final co-ordinate of the particle is

$$X' = r + X_1' + X_2' + \dots + X_n' = 2S_n + r - n$$

where $S_n = \sum X_i$.

From the conditions of the question, X_i' 's or X_i 's are mutually independent, and hence S_n is binomial (n, p) . Therefore, the spectrum of X' is given by (cf. Ex. 4 Sec 5.9)

$$\left. \begin{aligned} x_i' &= 2i + r - n & (i = 0, 1, 2, \dots, n) \\ P(X' = x_i') &= P(S_n = i) = \binom{n}{i} p^i q^{n-i} \end{aligned} \right\} \quad (8.9.1)$$

C. CONDITIONAL EXPECTATIONS AND REGRESSION

8.10 CONDITIONAL EXPECTATION

Discrete case. The *conditional expectation* or *mean value* of a continuous function $g(X, Y)$ of X and Y on the hypothesis $Y = y_j$ is defined by

$$E\{g(X, Y) | Y = y_j\} = \sum_i g(x_i, y_j) f_{i|j} = \frac{\sum_i g(x_i, y_j) f_{ij}}{f_{\cdot j}} \quad (8.10.1)$$

its existence being understood in the usual sense.

We note that $E\{g(X, Y) | Y = y_j\}$ is nothing but the expectation of the function $g(X, y_j)$ of X with respect to the conditional distribution of X on the hypothesis $Y = y_j$.

The *conditional mean* of X or the *mean of the conditional distribution* of X on the hypothesis $Y = y_j$ is defined by

$$m_{x|j} = E(X | Y = y_j) = \frac{\sum_i x_i f_{ij}}{f_{\cdot j}} \quad (8.10.2)$$

which obviously represents the centre of mass of the probability mass points on the line $y = y_j$.

The *conditional variance* of X on the hypothesis $Y = y_j$ is likewise given by

$$\sigma_{x|j}^2 = \text{var}(X | Y = y_j) = E\{(X - m_{x|j})^2 | Y = y_j\} \quad (8.10.3)$$

The definitions of other characteristics of the conditional distribution of X on the hypothesis $Y = y_j$ may be easily constructed. The conditional expectation of $g(X, Y)$ and the conditional mean, variance etc. of Y on the hypothesis $X = x_i$ are also defined in an exactly similar manner.

If X and Y are *independent*, we have

$$f_{i|j} = f_{xi}, \quad f_{j|i} = f_{yj}$$

and hence

$$E\{g(X) | Y = y_j\} = E\{g(X)\}, \quad E\{h(Y) | X = x_i\} = E\{h(Y)\} \quad (8.10.4)$$

If follows, in particular, that

$$m_{x|j} = m_x, \quad m_{y|i} = m_y, \quad \sigma_{x|j} = \sigma_x, \quad \sigma_{y|i} = \sigma_y \quad (8.10.5)$$

Continuous case. The *conditional expectation* of $g(X, Y)$ on the hypothesis $Y=y$ is defined by

$$\begin{aligned} E\{g(X, Y) | Y=y\} &= \int_{-\infty}^{\infty} g(x, y) f_x(x|y) dx \\ &= \frac{\int_{-\infty}^{\infty} g(x, y) f(x, y) dx}{f_y(y)} \end{aligned} \quad (8.10.6)$$

The *conditional mean* of X on the hypothesis $Y=y$ is a function of y to be denoted by $m_{x|y}$ or, more conveniently, by $m_x(y)$ and defined by

$$m_x(y) = E(X | Y=y) = \frac{\int_{-\infty}^{\infty} x f(x, y) dx}{f_y(y)} \quad (8.10.7)$$

Similarly, we define

$$m_y(x) = E(Y | X=x) = \frac{\int_{-\infty}^{\infty} y f(x, y) dy}{f_x(x)} \quad (8.10.8)$$

The conditional variances $\sigma_{x|y}^2$ or $\sigma_x^2(y)$ and $\sigma_{y|x}^2$ or $\sigma_y^2(x)$ are defined by

$$\sigma_x^2(y) = \text{var}(X | Y=y) = E[\{X - m_x(y)\}^2 | Y=y] \quad (8.10.9)$$

$$\sigma_y^2(x) = \text{var}(Y | X=x) = E[\{Y - m_y(x)\}^2 | X=x] \quad (8.10.10)$$

If X and Y are *independent*

$$f_x(x|y) = f_x(x), \quad f_y(y|x) = f_y(y)$$

which lead to

$$m_x(y) = m_x, \quad m_y(x) = m_y, \quad \sigma_x^2(y) = \sigma_x^2, \quad \sigma_y^2(x) = \sigma_y^2 \quad (8.10.11)$$

8.11 REGRESSION CURVES

In another terminology, the conditional mean $m_y(x)$, for a continuous distribution, is called the *regression function of Y on X* and the curve

$$y = m_y(x) \quad (8.11.1)$$

the regression curve of Y on X or sometimes the regression curve for the mean of Y . (Regression is a peculiar word which has come into use whose literal meaning has very little to do with its mathematical definition !) Geometrically, the regression function $m_y(x)$ represents the y -co-ordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by x and $x+dx$, which follows readily from (8.10.8), and hence the regression curve of Y on X is the locus of this centre of mass as x varies.

Similarly, the regression function of X on Y is $m_x(y)$, and the regression curve of X on Y is given by

$$x = m_x(y) \quad (8.11.2)$$

Thus equations (8.11.1) and (8.11.2) give the two regression curves of a continuous bivariate distribution.

In case a regression curve is a straight line, the corresponding regression is said to be *linear*. If one of the regressions is linear, it does not, however, follow that the other is also linear.

Remark. We can also develop the idea of regression curves for the mean for discrete distributions, but that is relatively unimportant. In the discrete case, the analogue of the regression curve of Y on X will not be a continuous curve but a disconnected set of points, viz. $(x_i, m_{y|i})$ ($i=0, \pm 1, \pm 2, \dots$) ; we may, if we like, connect the consecutive points by straight lines for convenience.

1. The expectation of the regression function of Y on X treated as a random variable, i.e. of $m_y(X)$ is readily obtained from (8.10.8) which gives

$$E\{m_y(X)\} = \int_{-\infty}^{\infty} m_y(x) f_x(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy$$

or

$$E\{m_y(X)\} = m_y \quad (8.11.3)$$

Hence

$$\sigma^2\{m_y(X)\} = E[\{m_y(X) - m_y\}^2] \quad (8.11.4)$$

This gives a measure of deviation of the regression curve $y = m_y(x)$ from the horizontal line $y = m_y$.

2. We know that the conditional variance $\sigma_y^2(x)$ is a measure of dispersion of the conditional distribution of Y on the hypothesis $X=x$, and hence that of the two-dimensional mass distribution lying in the strip between x and $x+dx$ about the conditional mean $m_y(x)$, for a fixed value of x . Let us now see if $E\{\sigma_y^2(X)\}$ also reduces to σ_y^2 or not.

$$E\{\sigma_y^2(X)\} = \int_{-\infty}^{\infty} \sigma_y^2(x) f_x(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{y - m_y(x)\}^2 f(x, y) dx dy$$

or

$$E\{\sigma_y^2(X)\} = E[\{Y - m_y(X)\}^2] \quad (8.11.5)$$

which is, in general, different from σ_y^2 and gives a measure of dispersion of the bivariate distribution about the regression curve $y = m_y(x)$. This is called the *variance of Y about the regression function of Y on X* and denoted by σ_{yx}^2 , i.e.

$$\sigma_{yx}^2 = E[\{Y - m_y(X)\}^2] \quad (8.11.6)$$

We define σ_{xy}^2 similarly.

3. **Minimum property.** An immediate consequence of the fact that the conditional second moment is minimum when taken about the corresponding conditional mean is the following important theorem: For any continuous function $g(x)$, $E[\{Y - g(X)\}^2]$ is minimum when $g(x) = m_y(x)$.

$$\begin{aligned} \text{Proof. } E[\{Y - g(X)\}^2] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{y - g(x)\}^2 f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} f_x(x) dx \int_{-\infty}^{\infty} \{y - g(x)\}^2 f_y(y|x) dy \\ &= \int_{-\infty}^{\infty} f_x(x) dx E[\{Y - g(x)\}^2 | X=x] \end{aligned}$$

Now the expectation within the integral represents the conditional second moment of Y on the hypothesis $X=x$ about the point $g(x)$,

which, we know, is minimum when $g(x) = m_y(x)$, and hence the theorem follows.

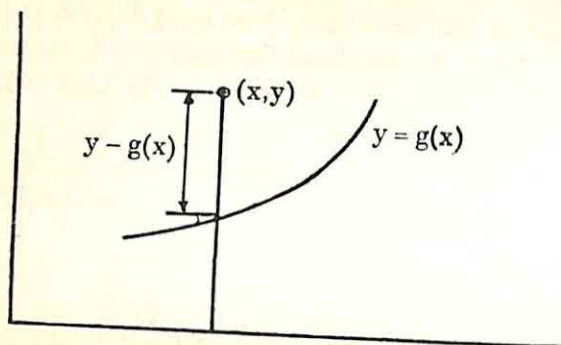


Fig. 18

The geometrical interpretation of the above theorem reveals the following minimum property of the regression curves. Here $y = g(x)$ represents any continuous curve, and $E[\{Y - g(X)\}^2]$ is the mean value of the square of the deviation of the distribution from the curve $y = g(x)$ measured in the direction of the y -axis, which is thus a measure of dispersion about the curve $y = g(x)$. Now the theorem states that among all continuous curves this mean value is minimum for the regression curve $y = m_y(x)$. Hence among all continuous curves, the one which minimises the mean value of the square of the deviation in the direction of the y -axis is the regression curve of Y on X .

4. If X and Y are independent, it follows from (8.10.11) that both the regressions : Y on X and X on Y are linear, the regression curves being $y = m_y$ and $x = m_x$ which are straight lines parallel to the axes.

Examples

1. Find the regression curves for the mean in the example of Sec. 6.3.

$$\int_{-\infty}^{\infty} y f(x, y) dx = 6 \int_0^{1-x} y(1-x-y) dx = (1-x)^3$$

So

$$m_y(x) = \frac{1}{3}(1-x) \quad (0 < x < 1)$$

Hence the regression curve for the mean of Y is

$$y = \frac{1}{3}(1 - x) \quad (0 < x < 1)$$

Similarly the regression curve for the mean of X is

$$x = \frac{1}{3}(1 - y) \quad (0 < y < 1)$$

2. BIVARIATE NORMAL DISTRIBUTION. From Ex. 3 Sec. 6.5 we have

$$m_y(x) = m_y + \rho \frac{\sigma_y}{\sigma_x} (x - m_x)$$

Hence the regression curve of Y on X is

$$y = m_y + \rho \frac{\sigma_y}{\sigma_x} (x - m_x)$$

or

$$\frac{y - m_y}{\sigma_y} = \rho \frac{x - m_x}{\sigma_x}$$

Similarly, the regression curve of X on Y is

$$\frac{y - m_y}{\sigma_y} = \frac{1}{\rho} \frac{x - m_x}{\sigma_x}$$

Thus for the bivariate normal distribution both the regressions are linear.

8.12 LEAST SQUARE REGRESSION CURVES

By dilating the minimum property only, we can introduce a very general and useful concept of regression curves known as *least square regression curves*. The principle of least square is, however, a broad mathematical principle which, in this case, may be precisely stated as follows. Let

$$y = g(x; c_0, c_1, \dots) \quad (8.12.1)$$

be a family of curves, c_0, c_1, \dots being the *parameters* of the family. The principle of least squares consists in minimising the mean value

$$S = E[\{Y - g(X; c_0, c_1, \dots)\}^2] \quad (8.12.2)$$

which is a function of the parameters c_0, c_1, \dots and which gives a measure of dispersion of the probability mass distribution about the

curve (8.12.1). If S is minimum for $c_0 = c_0^*$, $c_1 = c_1^*$,, then the curve

$$y = g(x; c_0^*, c_1^*, \dots) \quad (8.12.3)$$

is said to be the *best-fitting curve of the family* (8.12.1) to the distribution according to the principle of least squares and will be called the *least square regression curve of Y on X belonging to the given family*. The function $g(x; c_0^*, c_1^*, \dots)$ is called the *least square regression function of Y on X* , and the corresponding random variable

$$U_y = g(X; c_0^*, c_1^*, \dots) \quad (8.12.4)$$

the *best representation of Y by a function of X of the family* $g(X; c_0, c_1, \dots)$ according to the least square principle. The variate

$$V_y = Y - U_y \quad (8.12.5)$$

which is the part of Y left after taking away its best representation is called the *residual of Y* . We note

$$S_{min} = E\{[Y - g(X; c_0^*, c_1^*, \dots)]^2\} = E(V_y^2) \quad (8.12.6)$$

Clearly, S_{min} is a measure of dispersion about the regression curve (8.12.3) and hence an inverse measure of goodness of fit of the regression curve to the probability distribution.

The equations for minimising S are

$$\frac{\partial S}{\partial c_0} = 0, \quad \frac{\partial S}{\partial c_1} = 0, \dots \quad (8.12.7)$$

which are called the *normal equations*. By solving these equations we get the *least square values* c_0^* , c_1^* ,, of the parameters c_0, c_1, \dots respectively.

Similar formulations hold for the least square regression curves of X on Y . If, however, we consider the family of *all* continuous curves, it follows from the last section that for a continuous distribution the least square regression curve of Y on X turns out to be the regression curve for the mean of Y . Thus we see that the regression for the mean may be obtained as a particular case of least square regression.

The most important of the least square regression curves are the regression lines, although other types of curves are also sometimes used. We shall, for convenience, often omit the phrase *least square* qualifying a regression curve which will be implied by the context.

8.13 REGRESSION LINES

Here we are concerned with the family of straight lines

$$y = c_0 + c_1 x \quad (8.13.1)$$

so that

$$S = E\{(Y - c_0 - c_1 X)^2\} \quad (8.13.2)$$

The normal equations are $\frac{\partial S}{\partial c_0} = 0$ and $\frac{\partial S}{\partial c_1} = 0$ which, on putting $c_0 = c_0^*$, $c_1 = c_1^*$, reduce to

$$E(Y - c_0^* - c_1^* X) = 0 \quad (8.13.3)$$

and

$$E\{X(Y - c_0^* - c_1^* X)\} = 0 \quad (8.13.4)$$

or

$$\begin{aligned} c_0^* + c_1^* m_x &= m_y \\ c_0^* m_x + c_1^* a_{xx} &= a_{xy} \end{aligned}$$

Solving these we get

$$c_0^* = m_y - \rho \frac{\sigma_y}{\sigma_x} m_x, \quad c_1^* = \rho \frac{\sigma_y}{\sigma_x} \quad (8.13.5)$$

Therefore, the regression line of Y on X is

$$y = c_0^* + c_1^* x = m_y + \rho \frac{\sigma_y}{\sigma_x} (x - m_x) \quad (8.13.6)$$

The coefficient of x , $c_1^* = \rho \frac{\sigma_y}{\sigma_x}$ is called the *regression coefficient of Y on X* and denoted by β_{yx} , i.e.

$$\beta_{yx} = \rho \frac{\sigma_y}{\sigma_x} \quad (8.13.7)$$

Equation (8.13.6) may also be written in the form

$$\frac{y - m_y}{\sigma_y} = \rho \frac{x - m_x}{\sigma_x} \quad (8.13.8)$$

We have

$$\begin{aligned}(Y - c_0^* - c_1^* X)^2 &= \left\{ Y - m_y - \rho \frac{\sigma_y}{\sigma_x} (X - m_x) \right\}^2 \\ &= (Y - m_y)^2 + \rho^2 \frac{\sigma_y^2}{\sigma_x^2} (X - m_x)^2 - 2\rho \frac{\sigma_y}{\sigma_x} (X - m_x)(Y - m_y)\end{aligned}$$

So

$$E\{(Y - c_0^* - c_1^* X)^2\} = \sigma_y^2 + \rho^2 \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 - 2\rho \frac{\sigma_y}{\sigma_x} \rho \sigma_x \sigma_y$$

or

$$E\{(Y - c_0^* - c_1^* X)^2\} = \sigma_y^2 (1 - \rho^2) \quad (8.13.9)$$

Similarly, the regression line of X on Y is

$$x = d_0^* + d_1^* y$$

where

$$d_0^* = m_x - \rho \frac{\sigma_x}{\sigma_y} m_y, \quad d_1^* = \rho \frac{\sigma_x}{\sigma_y} \quad (8.13.10)$$

or

$$\frac{y - m_y}{\sigma_y} = \frac{1}{\rho} \frac{x - m_x}{\sigma_x} \quad (8.13.11)$$

The regression coefficient of X on Y ,

$$\beta_{xy} = \rho \frac{\sigma_x}{\sigma_y} \quad (8.13.12)$$

and

$$E\{(X - d_0^* - d_1^* Y)^2\} = \sigma_x^2 (1 - \rho^2) \quad (8.13.13)$$

1. **Significance of ρ .** We remarked earlier that if $\rho = 0$, X and Y are not necessarily independent, but if $\rho = \pm 1$, we can conclude that Y is a linear function of X . The latter result becomes obvious from (8.13.9) or (8.13.13). If $\rho = \pm 1$, the regression lines (8.13.8) and (8.13.11) coincide, and it follows that

$$Y = c_0^* + c_1^* X = m_y \pm \frac{\sigma_y}{\sigma_x} (X - m_x)$$

which is a linear function of X , i.e. the whole probability mass is confined on the coincident regression lines. We shall now show that $|\rho|$, in fact, gives a *measure of linear dependence of X and Y* .

From (8.13.9) or (8.13.13) we get $0 \leq |\rho| \leq 1$, and the left-hand sides of (8.13.9) and (8.13.13), which are measures of dispersion about the corresponding regression lines, are both proportional to $1 - \rho^2$. This shows that $|\rho|$ is a measure of concentration of the probability mass about the regression lines which are the best-fitting lines to the distribution, or, in other words, $|\rho|$ is a measure of linear dependence of X and Y . We may also say that $|\rho|$ is a *measure of goodness of fit* of the regression lines to the distribution. Moreover, it is a very satisfactory measure in view of the fact that the correlation coefficient is dimensionless and independent of the choice of origins and scales of measurements.

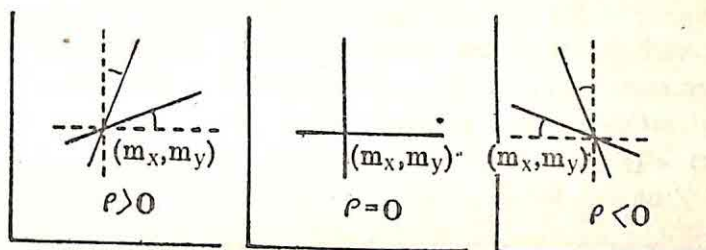


Fig. 19. Least Square Regression Lines

2. When $\rho = 0$, the regression lines (8.13.8) and (8.13.11) respectively reduce to $y = m_y$ and $x = m_x$. For $\rho > 0$, both the regression lines have positive slopes, while for $\rho < 0$ they have negative slopes.

3. The best representation of Y ,

$$U_y = c_0^* + c_1^* X = m_y + \rho \frac{\sigma_y}{\sigma_x} (X - m_x) \quad (8.13.14)$$

Hence

$$E(U_y) = m_y, \quad \sigma(U_y) = |\rho| \sigma_y \quad (8.13.15)$$

and

$$\rho(U_y, Y) = |\rho| \geq 0 \quad (8.13.16)$$

Thus we may say that the correlation coefficient between Y and its best representation is a measure of goodness of fit of the regression lines to the probability distribution.

4. The residual

$$V_y = Y - U_y = Y - c_0^* - c_1^* X \quad (8.13.17)$$

Now the normal equations state

$$E(V_y) = 0, \quad E(XV_y) = 0 \quad (8.13.18)$$

By (8.13.9)

$$\sigma^2(V_y) = E(V_y^2) = \sigma_y^2(1 - \rho^2) \quad (8.13.19)$$

or using (8.13.15)

$$\sigma^2(V_y) = \sigma_y^2 - \sigma^2(U_y) \quad (8.13.20)$$

This expresses that the variance of the residual of Y is the amount of the variance of Y left after subtraction of the variance of its best representation; in this sense $\sigma^2(V_y)$ is sometimes called the *residual variance of Y* . It follows from (8.13.19) that the residual variance $\sigma^2(V_y)$ is a measure of dispersion about the regression line of Y on X .

From (8.13.18) $\text{cov}(X, V_y) = E\{(X - m_x)V_y\} = 0$ so that

$$\rho(X, V_y) = 0 \quad (8.13.21)$$

Also then

$$\rho(U_y, V_y) = 0 \quad (8.13.22)$$

5. If the regression curve for the mean of Y , $y = m_y(x)$ happens to be a straight line, then it must be identical with the least square regression line of Y on X , for while competing among all possible continuous curves, the straight line $y = m_y(x)$ is the minimising curve, and hence it must also be the minimising curve while competing in its own family of straight lines.

Remark. The idea of least square fitting in the case of a one-dimensional distribution, although trivial, will be interesting to note. The problem here will reduce to fitting a point c to the distribution of X by minimising $E\{(X - c)^2\}$; the normal equation is $E(X - c) = 0$ and, therefore, $c^* = E(X) = m$. Thus the mean of any distribution is best-fitting point to the distribution according to the principle of least squares.

Examples

1. Find the regression lines in Ex. 1 Sec. 6.2. From Ex. 1 Sec. 8.3.

$$\beta_{yx} = \mu_{11}/\sigma_x^2 = -\frac{1}{2}, \quad \beta_{xy} = \mu_{11}/\sigma_y^2 = -\frac{5}{16}$$

Hence the regression lines are

$$y - \frac{10}{9} = -\frac{1}{2}(x - \frac{7}{9}) \quad (Y \text{ on } X)$$

$$x - \frac{7}{9} = -\frac{5}{16}(y - \frac{10}{9}) \quad (X \text{ on } Y)$$

2. Find the regression lines in the example of Sec. 6.3.

Here

$$m_x = m_y = \frac{1}{2}, \quad \sigma_x^2 = \sigma_y^2 = \frac{3}{8}, \quad \rho = -\frac{1}{3}$$

so that $\beta_{yx} = \beta_{xy} = -\frac{1}{3}$. Hence the regression line of Y on X and that of X on Y are respectively

$$y - \frac{1}{2} = -\frac{1}{3}(x - \frac{1}{2}), \quad x - \frac{1}{2} = -\frac{1}{3}(y - \frac{1}{2})$$

or

$$y = \frac{1}{3}(1 - x), \quad x = \frac{1}{3}(1 - y)$$

It appears in Ex. 1 Sec. 8.11 that these regressions lines are the same as the corresponding regression curves for the means. In fact, in this case both the regression curves for the means are straight lines and hence must coincide with the regression lines.

8.14 PARABOLIC CURVE FITTING

For the regression of Y on X , we consider the family of k th degree parabolas

$$y = c_0 + c_1x + c_2x^2 + \dots + c_kx^k \quad (8.14.1)$$

c_0, c_1, \dots, c_k being the $(k+1)$ parameters of the family. Set

$$S = E\{(Y - c_0 - c_1X - c_2X^2 - \dots - c_kX^k)^2\} \quad (8.14.2)$$

The normal equations are

$$\frac{\partial S}{\partial c_0} = 0, \quad \frac{\partial S}{\partial c_1} = 0, \quad \dots, \quad \frac{\partial S}{\partial c_k} = 0$$

Hence, for $c_0 = c_0^*, c_1 = c_1^*, \dots, c_k = c_k^*$, we have

$$\begin{aligned} E\{(Y - c_0^* - c_1^*X - c_2^*X^2 - \dots - c_k^*X^k)\} &= 0 \\ E\{X(Y - c_0^* - c_1^*X - c_2^*X^2 - \dots - c_k^*X^k)\} &= 0 \\ \dots &\dots \dots \\ E\{X^k(Y - c_0^* - c_1^*X - c_2^*X^2 - \dots - c_k^*X^k)\} &= 0 \end{aligned} \quad (8.14.3)$$

In terms of the moments a_{kl} , these reduce to

$$\begin{aligned} c_0^* a_{00} + c_1^* a_{10} + c_2^* a_{20} + \dots + c_k^* a_{k0} &= a_{01} \\ c_0^* a_{10} + c_1^* a_{20} + c_2^* a_{30} + \dots + c_k^* a_{k+1,0} &= a_{11} \\ &\dots \dots \dots \\ c_0^* a_{k0} + c_1^* a_{k+1,0} + c_2^* a_{k+2,0} + \dots + c_k^* a_{2k,0} &= a_{k1} \end{aligned} \quad (8.14.4)$$

These equations give $c_0^*, c_1^*, \dots, c_k^*$. The best-fitting parabola of degree k is then

$$y = c_0^* + c_1^* X + c_2^* X^2 + \dots + c_k^* X^k \quad (8.14.5)$$

and the best representation of Y by a k th degree polynomial in X is

$$U_y = c_0^* + c_1^* X + c_2^* X^2 + \dots + c_k^* X^k \quad (8.14.6)$$

The residual is given by

$$V_y = Y - U_y = Y - c_0^* - c_1^* X - \dots - c_k^* X^k \quad (8.14.7)$$

The first normal equation states

$$E(V_y) = 0 \quad (8.14.8)$$

Hence

$$\sigma^2(V_y) = E(V_y^2) = S_{min} \quad (8.14.9)$$

Now we can easily calculate S_{min} by the following tricky use of the normal equations. Write

$$S = E\{(kY - c_0 - c_1 X - \dots - c_k X^k)^2\}$$

where $k=1$, so that S becomes a homogeneous function of k, c_0, c_1, \dots, c_k of degree 2. Hence

$$2S = k \frac{\partial S}{\partial k} + c_0 \frac{\partial S}{\partial c_0} + c_1 \frac{\partial S}{\partial c_1} + \dots + c_k \frac{\partial S}{\partial c_k}$$

and

$$\begin{aligned} 2S_{min} &= \left[k \frac{\partial S}{\partial k} + c_0 \frac{\partial S}{\partial c_0} + c_1 \frac{\partial S}{\partial c_1} + \dots \right. \\ &\quad \left. + c_k \frac{\partial S}{\partial c_k} \right]_{k=1, c_0=c_0^*, \dots, c_k=c_k^*} \\ &= \frac{\partial S}{\partial k} \Big|_{k=1, c_0=c_0^*, \dots, c_k=c_k^*} \\ &= 2E\{Y(Y - c_0^* - c_1^* X - \dots - c_k^* X^k)\} \end{aligned}$$

or

$$S_{min} = a_{02} - c_0^* a_{01} - c_1^* a_{11} - \dots - c_k^* a_{k1} \quad (8.14.10)$$

We know that S_{min} may be taken to be an inverse measure of goodness of fit of the regression parabola (8.14.5) to the probability distribution; but this measure has the obvious defect of being not dimensionless, and also its range of variation is unknown. For the regression lines, however, it was found that the numerical value of the correlation coefficient furnishes a dimensionless measure of fit, and that $0 \leq |\rho| \leq 1$. In order to obtain such a satisfactory measure for polynomial regression also, we first reduce it to a case of linear regression by the transformation

$$U_y = c_0^* + c_1^* X + \dots + c_k^* X^k \quad (8.14.11)$$

Consider now the joint distribution of the variates U_y and Y which is uniquely determined by the given bivariate distribution of X , Y , and let us find the regression line of Y on U_y . For this, we consider family of straight lines

$$y = a + bu_y$$

in the (u_y, y) -plane and minimise

$$E\{(Y - a - bU_y)^2\} = E\{[Y - (a + bc_0^*) - bc_1^* X - \dots - bc_k^* X^k]^2\}$$

Since the R.H.S. is of the form (8.14.2), it is obviously minimum for $a = 0$, $b = 1$, so that the regression line of Y on U_y is $y = u_y$.

Therefore, the best representation of Y by a linear function of U_y is U_y , and the corresponding residual of Y is $Y - U_y = V_y$.

It follows from (8.13.16) that $\rho(U_y, Y) \geq 0$, and hence

$$0 \leq \rho(U_y, Y) \leq 1 \quad (8.14.12)$$

By (8.13.19)

$$\sigma^2(V_y) = \sigma_y^2 \{1 - \rho^2(U_y, Y)\} \quad (8.14.13)$$

which at once shows that the correlation coefficient $\rho(U_y, Y)$, whose limits are given by (8.14.12), is the required dimensionless measure of goodness of fit of the regression parabola of Y on X to the distribution. If also follows from (8.13.15), (8.13.20), (8.12.22) that

$$E(U_y) = m_y, \quad \sigma(U_y) = \rho(U_y, Y) \sigma_y \quad (8.14.14)$$

$$\sigma^2(V_y) = \sigma_y^2 - \sigma^2(U_y), \quad \rho(U_y, V_y) = 0$$

where U_y and V_y are given by (8.14.6) and (8.14.7) respectively.

8.15 CORRELATION RATIO

Returning to the topic of regression for the mean, we may also be interested in constructing a dimensionless measure of goodness of fit of the regression curve, say, for the mean of Y , $y = m_y(x)$. We know that the problem of regression for the mean can also be considered as a least square regression problem corresponding to the family of all possible continuous curves, and, by following a method similar to that in the last section, we reduce this problem to one of least square regression lines by setting

$$U_y' = m_y(X) \quad (8.15.1)$$

which is evidently the best representation of Y by any continuous function of X .

It is easy to see that, for the joint distribution of U_y' and Y , the least square regression line of Y on U_y' is $y = u_y'$, for

$$E\{(Y - a - bU_y')^2\} = E\{[Y - a - bm_y(X)]^2\}$$

is certainly minimum when $a = 0$, $b = 1$.

Hence the best representation of Y by a linear function of U_y' is simply U_y' , and the residual of Y is $Y - U_y' = V_y'$ (say). We note that V_y' may also be interpreted as the residual of Y corresponding to its best representation by any continuous function of X .

From (8.13.18) $E(V_y') = 0$, and hence by (8.11.6)

$$\sigma^2(V_y') = E(V_y'^2) = E\{[Y - m_y(X)]^2\} = \sigma_{yx}^2$$

By (8.13.16) and (8.13.19)

$$0 \leq \rho\{m_y(X), Y\} \leq 1 \quad (8.15.2)$$

and

$$E\{[Y - m_y(X)]^2\} = \sigma_y^2[1 - \rho^2\{m_y(X), Y\}] \quad (8.15.3)$$

This shows that $\rho\{m_y(X), Y\}$ is a measure of goodness of fit of the regression curve $y = m_y(x)$ to the distribution, which is dimensionless and non-negative. This correlation coefficient between the regression function of Y on X (treated as a random variable) and Y is called the *correlation ratio of Y on X* and denoted by η_{yx} , i.e.

$$\eta_{yx} = \rho\{m_y(X), Y\} \quad (8.15.4)$$

By (8.15.2) and (8.15.3)

$$0 \leq \eta_{yx} \leq 1 \quad (8.15.5)$$

$$\sigma_{yx}^2 = \sigma_y^2 (1 - \eta_{yx}) \quad (8.15.6)$$

Also (8.13.15), (8.13.20), (8.13.22) will hold if we replace U_y and V_y by U_y' and V_y' respectively, i.e.

$$E\{m_y(X)\} = m_y, \quad \sigma\{m_y(X)\} = \eta_{yx} \sigma_y \quad (8.15.7)$$

$$\sigma_{yx}^2 = \sigma_y^2 - \sigma^2\{m_y(X)\}, \quad \rho\{m_y(X), Y - m_y(X)\} = 0$$

1. If $\eta_{yx} = 1$, $\sigma_{yx} = 0$, and hence $Y = m_y(X)$, i.e. all the probability mass is situated on the regression curve $y = m_y(x)$.

2. If $\eta_{yx} = 0$, $\sigma\{m_y(X)\} = 0$, hence $m_y(x) = m_y$, i.e. the regression curve for the mean of Y is the straight line $y = m_y$, which must then be identical with the regression line (8.13.8). Comparing these two we get $\beta_{yx} = 0$ or $\rho = 0$. Thus $\eta_{yx} = 0$ implies $\rho = 0$.

3. In case the regression function $m_y(x)$ is linear, we must have $m_y(x) = c_0 + c_1 x$, so that from (8.15.4) $\eta_{yx} = |\rho|$.

Example. Calculate σ_{yx}^2 and η_{yx} in the example of Sec. 6.3.

From Ex. 1 Sec. 8.11

$$\sigma_{yx}^2 = 6 \int_0^1 \int_0^{1-x} [y - \frac{1}{3}(1-x)]^2 (1-x-y) dx dy = \frac{1}{80}$$

Since $\sigma_y^2 = \frac{8}{80}$, by (8.15.6) $\eta_{yx} = \frac{1}{3}$.

In fact, here the regression function $m_y(x)$ is linear, and so by observation 3 above $\eta_{yx} = |\rho| = \frac{1}{3}$.

8.16 EXERCISES

1. Two points X, Y are independently chosen at random on a line segment of length a . Compute the expectation of $|X - Y|^n$.

2. If (X, Y) has the normal distribution in two dimensions with zero means, unit variances and correlation coefficient ρ , then prove that the expectation of the greater of X and Y is $\sqrt{(1-\rho)/\pi}$.

3. If X_1, X_2, \dots, X_n are mutually independent standard normal variates, then show that the mean value of $\min(|X_1|, |X_2|, \dots, |X_n|)$ is

$$2^n \int_0^{\infty} [1 - \Phi(x)]^n dx$$

where $\Phi(x)$ denotes the standard normal distribution function.

4. A ball is drawn at random from an urn containing 3 white balls numbered 0, 1, 2, 2 red balls numbered 0, 1 and 1 black ball numbered 0. If the colours white, red and black are again numbered 0, 1 and 2 respectively, find the correlation coefficient between the variates— X , the colour number and Y , the number of the ball. Also write down the least square regression lines of Y on X and X on Y .

5. In Ex. 1 Sec. 6.8 find the regression lines of the joint distribution of the random variables—the number of the ball and the colour number, and obtain a measure of goodness of fit.

6. Let X and Y respectively denote the number of heads and the longest run of heads in four tosses of a coin. Compute the means, variances and the correlation coefficient.

7. When two dice are thrown, X denotes the number on the first die and Y the greater of the two numbers on the dice. Compute the correlation coefficient between X and Y .

8. Calculate m_x , m_y , and a_{11} for the joint distribution of the number of the ball and that of the colour in Ex. 2 Sec. 6.8. Hence find the covariance of the variates, and explain the value obtained.

9. A ball is drawn from an urn containing 3 balls marked with numbers 0, 1, 2 and having white, red and black colours respectively. If we call white, red and black the first, second and third colour respectively, show that the correlation coefficient between X , the number of the colour and Y , that of the ball is unity, and hence obtain Y as a function of X .

10. The probability density function of a continuous bivariate distribution is given by

$$\begin{aligned} f(x, y) &= x+y && \text{for } 0 < x < 1, 0 < y < 1 \\ &= 0 && \text{elsewhere} \end{aligned}$$

Find the values of m_x , m_y , σ_x , σ_y , ρ and write down the regression lines. Also find the regression curves for the means.

11. For the continuous bivariate distribution defined in Ex. 8. Sec. 6.8, find the regression curves for the means, and also the the least square regression lines.

12. Show that the acute angle θ between the least square regression lines is given by

$$\tan \theta = \frac{1-\rho^2}{\rho} \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

and discuss the cases $\rho=0$ and $\rho=\pm 1$.

13. If the regression lines are

$$x+6y=6 \quad \text{and} \quad 3x+2y=10$$

find the means and the correlation coefficient.

14. If points are assigned to the different suits of cards as follows: 1 for spade, 2 for heart, 3 for diamond and 4 for club, find the mathematical expectation of the total number of points in a bridge hand of 13 cards.

15. Two points are independently chosen at random on two adjacent sides of a square, the length of a side being a . Find the mean area of the triangle formed by the line joining the two random points and the sides of the square.

16. Prove Schwartz's inequality for expectations that $[E(XY)]^2 \leq E(X^2)E(Y^2)$, and hence deduce that $-1 \leq \rho(X, Y) \leq 1$.

17. If a, b, c are positive constants, show that the correlation coefficient between $aX+bY$ and cY is

$$\frac{a\rho\sigma_x + b\sigma_y}{\sqrt{a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y}}$$

where $\rho = \rho(X, Y)$.

18. If X and Y are uncorrelated, find the correlation coefficient between the linear combinations a_1X+b_1Y and a_2X+b_2Y .

19. Prove that the linear combinations $aX+bY$ and $cX+dY$ of the random variables X, Y are uncorrelated if $ac\sigma_x^2 + (ad+bc)\rho\sigma_x\sigma_y + bd\sigma_y^2 = 0$.

20. The random variables X, Y are connected by the linear relation $aX+bY+c=0$. Prove that the correlation coefficient between X and Y is -1 if a, b have the same sign and 1 if a, b have opposite signs.

21. In the theorem of Sec. 8.3 show that if U, V are uncorrelated, then $\sigma_u^2 + \sigma_v^2 = \sigma_x^2 + \sigma_y^2$, $\sigma_u\sigma_v = \sigma_x\sigma_y\sqrt{1-\rho^2}$.

22. If for any pair of linearly dependent random variables X, Y we set

$$U = X \cos \alpha + Y \sin \alpha, \quad V = -X \sin \alpha + Y \cos \alpha \quad (\alpha, \text{ a constant})$$

then prove that V will be constant (i.e. has a one-point distribution) if $\tan \alpha = \rho\sigma_y/\sigma_x$.

23. If the joint distribution of X and Y is the bivariate normal distribution, then show that

$$\frac{X}{\sigma_x} + \frac{Y}{\sigma_y} \quad \text{and} \quad \frac{X}{\sigma_x} - \frac{Y}{\sigma_y}$$

are independent normal variates having variances $2(1+\rho)$ and $2(1-\rho)$ respectively where $\rho = \rho(X, Y)$.

24. Show that the mean and variance of the number of successes in a Poisson sequence of n trials are $\sum p_i$ and $\sum p_i q_i$ respectively, where p_i denotes the probability of success in the i th trial and $q_i = 1 - p_i$ ($i = 1, 2, \dots, n$).

25. Find the mean and variance of the total number of aces, kings, queens and jacks obtained by a bridge hand.

26. An urn contains N tickets numbered $1, 2, \dots, N$, from which n tickets are drawn successively without replacement. Find the mean and variance of the sum of the numbers on the tickets drawn.

27. In Ex. 5 Sec. 3.2 show that the mean and variance of the number of matches are both unity.

28. If the mutually independent random variables X_1, X_2, \dots, X_n all have the same distribution and their sum $X_1 + X_2 + \dots + X_n$ is normally distributed, then show that each of them is normally distributed.

29. Let X_1, X_2, \dots, X_n be mutually independent random variables having cumulants $\kappa_k^{(1)}, \kappa_k^{(2)}, \dots, \kappa_k^{(n)}$ respectively. Prove that the linear combination $X = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$ has cumulants κ_k given by

$$\kappa_k = a_1^k \kappa_k^{(1)} + a_2^k \kappa_k^{(2)} + \dots + a_n^k \kappa_k^{(n)}$$

Hence deduce that the mean m and variance σ^2 of X are given by

$$m = a_1 m_1 + a_2 m_2 + \dots + a_n m_n, \quad \sigma^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2$$

where m_i and σ_i^2 respectively denote the mean and variance of X_i ($i = 1, 2, \dots, n$).

30. A gambler plays a game of chance with his opponent, in which the probability of his win is $\frac{1}{2}$ at a stake of 1 rupee. If he starts with a capital of 5 rupees, find the probability that the gambler will just completely ruin his capital after 9 games. (Assume that if the capital of the gambler is exhausted before 9 games are played, he can be given loan so that the play is not stopped.)

31. The joint probability distribution of two discrete random variables X, Y is given by

$$P(X=i, Y=j) = p_{ij} \quad (i, j = 0, 1)$$

Find the joint characteristic function of X and Y and their individual characteristic functions, and using these prove that X, Y are independent if

$$p_{00}p_{11} = p_{01}p_{10}$$

32. Prove the formula

$$\sigma^2(V_y) = E(XV_y)$$

for the least square linear regression of Y on X . Deduce that

$$\sigma^2(V_y) = a_{02} - c_0^* a_{01} - c_1^* a_{11}$$

33. Fit a parabola of the form $y=c_0+c_1x+c_2x^2$ to the joint distribution of X and Y defined in Ex. 4 by the principle of least squares, and find a measure of goodness of fit.

34. Find the least square regression parabola of the second degree of Y on X for the bivariate normal distribution with zero means, and account for your result.

35. The joint probability density function of X and Y is $\frac{1}{2}x^2e^{-x(y+1)}$ ($0 < x < \infty$, $0 < y < \infty$). Determine the correlation ratio of Y on X .

36. For any continuous function $g(X)$, prove that

$$\rho\{g(X), Y - m_y(X)\} = 0$$

Use this fact to show that

$$E\{[m_y(X) - c_0 - c_1X]^2\} = \sigma_y^2(\eta_{yx}^2 - \rho^2)$$

Hence deduce that $\eta_{yx} \geq |\rho|$ and that $\eta_{yx}^2 - \rho^2$ is a measure of separation between the regression curve for the mean of Y and the corresponding regression line. Interpret the case $\eta_{yx} = |\rho|$.

SPECIAL DISTRIBUTIONS

In this chapter we shall study three continuous distributions, viz. the χ^2 , t and F -distributions which are particularly important for their applications in statistics. Now it is customary in statistics to denote the random variable associated with these distributions by the same letter as the distribution itself. We shall also initiate this practice here, and speak, for example, of a random variable χ^2 having a χ^2 -distribution, the corresponding running real variable being again denoted by x^2 in keeping with our usual convention.

9.1 χ^2 -DISTRIBUTION

The spectrum consists of the positive half of the real axis, and the density function is defined by

$$f(x^2) = \frac{e^{-\frac{1}{2}x^2} (\frac{1}{2}x^2)^{n/2-1}}{2\Gamma(\frac{1}{2}n)} \quad \begin{matrix} x^2 > 0 \\ = 0 & x^2 < 0 \end{matrix} \quad (9.1.1)$$

n , the only parameter of the distribution, is a positive integer called the *number of degrees of freedom* of the distribution. A χ^2 -distribution with n degrees of freedom will be briefly referred to as a $\chi^2(n)$ distribution.

The above form of the density function reminds one of the gamma distribution, and, in fact, we have the following theorem.

Theorem I. If X is a $\gamma(\frac{1}{2}n)$ variate, then $Y = 2X$ has a χ^2 -distribution with n degrees of freedom, and conversely if Y is a $\chi^2(n)$ variate, then X is a $\gamma(\frac{1}{2}n)$ variate.

Proof. The probability differential

$$dF = \frac{e^{-x} x^{n/2-1}}{\Gamma(\frac{1}{2}n)} dx = \frac{e^{-y/2} (\frac{1}{2}y)^{n/2-1}}{\Gamma(\frac{1}{2}n)} d(\frac{1}{2}y) = \frac{e^{-y/2} (\frac{1}{2}y)^{n/2-1}}{2\Gamma(\frac{1}{2}n)} dy \quad (0 < y < \infty)$$

which shows that Y has a $\chi^2(n)$ distribution. The proof of the converse will be similar.

The next theorem shows how the χ^2 -distribution arises naturally from the normal distribution.

Theorem II. If X_1, X_2, \dots, X_n are n mutually independent standard normal variates, then the sum of their squares $X_1^2 + X_2^2 + \dots + X_n^2$ is χ^2 -distributed with n degrees of freedom.

Proof. Since each X_i is normal $(0, 1)$, $\frac{1}{\sqrt{2}}X_i^2$ is a $\gamma(\frac{1}{2})$ variate (cf. Ex. 3 Sec. 5.9). Now $\frac{1}{\sqrt{2}}X_1^2, \frac{1}{\sqrt{2}}X_2^2, \dots, \frac{1}{\sqrt{2}}X_n^2$ are mutually independent γ -variates each with parameter $\frac{1}{2}$, so that, by the reproductive property of the γ -distribution, their sum $\frac{1}{2}(X_1^2 + X_2^2 + \dots + X_n^2)$ is a $\gamma(\frac{1}{2}n)$ variate, and hence, by Theorem I, $X_1^2 + X_2^2 + \dots + X_n^2$ is χ^2 -distributed with n degrees of freedom.

Theorem III. Let X_1, X_2, \dots, X_n be n mutually independent standard normal variates. If

$$\begin{array}{ccccccc} a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n \\ a_{21}X_1 + a_{22}X_2 + \dots + a_{2n}X_n \\ \dots & \dots & \dots \\ a_{m1}X_1 + a_{m2}X_2 + \dots + a_{mn}X_n \end{array} \quad (9.1.2)$$

be $m(<n)$ given linear combinations such that their coefficients satisfy the orthogonality relations :

$$\sum_{\alpha=1}^n a_{i\alpha} a_{j\alpha} = \delta_{ij} \quad (i, j = 1, 2, \dots, m) \quad (9.1.3)$$

where δ_{ij} is the well-known Kronecker symbol, i.e.

$$\delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$$

then the quadratic form $Q = Q(X_1, X_2, \dots, X_n)$ given by

$$Q = \sum_{\alpha=1}^n X_{\alpha}^2 - \sum_{\beta=1}^m (a_{\beta 1}X_1 + a_{\beta 2}X_2 + \dots + a_{\beta n}X_n)^2 \quad (9.1.4)$$

is χ^2 -distributed with $n - m$ degrees of freedom, and Q is independent of the given linear combinations.

Let us first prove the following lemma.

LEMMA. If X_1, X_2, \dots, X_n are mutually independent standard normal variates, and Y_1, Y_2, \dots, Y_n are obtained by an orthogonal homogeneous linear transformation :

$$Y_i = \sum_{\alpha=1}^n a_{i\alpha} X_{\alpha} \quad (i=1, 2, \dots, n) \quad (9.1.5)$$

where

$$\sum_{\alpha=1}^n a_{i\alpha} a_{j\alpha} = \sum_{\alpha=1}^n a_{\alpha i} a_{\alpha j} = \delta_{ij} \quad (i, j=1, 2, \dots, n) \quad (9.1.6)$$

then Y_1, Y_2, \dots, Y_n are also mutually independent standard normal variates.

Proof. Set

$$y_i = \sum_{\alpha=1}^n a_{i\alpha} x_{\alpha} \quad (i=1, 2, \dots, n)$$

It follows immediately from the orthogonality relations that

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n x_i^2, \quad \left| \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(y_1, y_2, \dots, y_n)} \right| = 1$$

Since X_1, X_2, \dots, X_n are mutually independent each normal $(0, 1)$, the probability differential

$$\begin{aligned} dF &= \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum x_i^2} dx_1 dx_2 \dots dx_n \\ &= \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum x_i^2} \left| \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(y_1, y_2, \dots, y_n)} \right| dy_1 dy_2 \dots dy_n \\ &= \frac{1}{(\sqrt{2\pi})^n} e^{-\sum y_i^2} dy_1 dy_2 \dots dy_n \\ &= \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \frac{1}{\sqrt{2\pi}} e^{-y_2^2/2} \dots \frac{1}{\sqrt{2\pi}} e^{-y_n^2/2} dy_1 dy_2 \dots dy_n \end{aligned}$$

This from of the probability differential proves the lemma. *

Proof of the theorem. We know from the theory of matrices that given $m \times n$ coefficients a_{ij} ($i=1, 2, \dots, m$; $j=1, 2, \dots, n$) of (9.1.2) subject to (9.1.3), we can always determine the rest of the n^2 numbers

a_{ij} ($i, j = 1, 2, \dots, n$) such that the complete set of orthogonality relations (9.1.6) holds giving an orthogonal transformation (9.1.5). Of these Y 's, we note, Y_1, Y_2, \dots, Y_m are the given linear combinations of (9.1.2).

$$\text{Now } \sum_{\alpha=1}^n X_{\alpha}^2 = \sum_{\alpha=1}^n Y_{\alpha}^2, \text{ and hence}$$

$$Q = \sum_{\alpha=1}^n Y_{\alpha}^2 - \sum_{\beta=1}^m Y_{\beta}^2 = Y_{m+1}^2 + Y_{m+2}^2 + \dots + Y_n^2$$

Since, by the lemma, Y_1, Y_2, \dots, Y_n are mutually independent standard normal variates, it follows from Theorem III (a) Sec. 6.7 that $Y_{m+1}, Y_{m+2}, \dots, Y_n$ are $n-m$ mutually independent standard normal variates, whence by Theorem II Q is a $\chi^2(n-m)$ variate. Again since Y_1, Y_2, \dots, Y_n are mutually independent, by Theorem III (c) Sec. 6.7, Q , a function of $Y_{m+1}, Y_{m+2}, \dots, Y_n$, must be independent of Y_1, Y_2, \dots, Y_m , i.e. of the given linear combinations.

Example. Let the Cartesian co-ordinates (X, Y, Z) of a random point in space be mutually independent, each of which is normal $(0, 1)$. By Theorem II, the square of the distance of the random point from the origin, $X^2 + Y^2 + Z^2$ is χ^2 -distributed with 3 degrees of freedom. Now consider any plane $lx + my + nz = 0$ ($l^2 + m^2 + n^2 = 1$) passing through the origin. The square of the distance from the origin of the foot of the perpendicular from the random point to the plane is $X^2 + Y^2 + Z^2 - (lX + mY + nZ)^2$ which, by Theorem III, is χ^2 -distributed with 2 degrees of freedom. This example might give a glimpse into the meaning of the term 'degrees of freedom'.

Characteristics of the χ^2 -distribution

We know, if X has a $\gamma(\frac{1}{2}n)$ distribution, $\chi^2 = 2X$ has a χ^2 -distribution with n degrees of freedom. Hence the moments of the $\chi^2(n)$ distribution are given by

$$a_k = 2^k a_k(X) = 2^{k\frac{1}{2}} n(\frac{1}{2}n + 1)(\frac{1}{2}n + 2) \dots (\frac{1}{2}n + k - 1)$$

or

$$a_k = n(n+2)(n+4) \dots (n+2k-2) \quad (9.1.7)$$

Hence mean $m = a_1 = n$, $a_2 = n(n+2)$ and variance $\sigma^2 = n(n+2) - n^2 = 2n$ etc.

For $n \leq 2$, the density function $f(\chi^2)$ is monotonic decreasing in $0 < x < \infty$, and as such the distribution has no mode; but for $n > 2$,

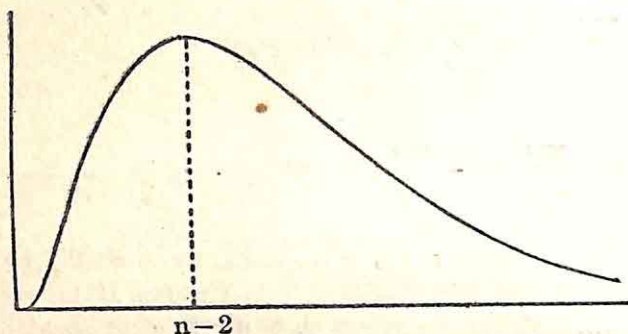


Fig. 20. Typical χ^2 -Density Curve

$f(\chi^2)$ has a single maximum at $\chi^2 = n-2$, i.e. the distribution is unimodal with mode $M = n-2$.

The characteristic function

$$\chi(t) = \chi_x(2t) = (1 - 2it)^{-n/2} \quad (9.1.8)$$

This form of the characteristic function at once suggests the following reproductive property of the χ^2 -distribution. If $\chi_1^2, \chi_2^2, \dots, \chi_k^2$ are mutually independent χ^2 -variates with degrees of freedom n_1, n_2, \dots, n_k respectively, then their sum $\chi_1^2 + \chi_2^2 + \dots + \chi_k^2$ is also χ^2 -distributed with $n_1 + n_2 + \dots + n_k$ degrees of freedom.

9.2 t-DISTRIBUTION

The *t-distribution* or *Student's distribution* ('Student' was the pseudonym of a statistician W. S. Gosset) is given by

$$f(t) = \frac{1}{\sqrt{n} B(\frac{1}{2}, \frac{1}{2}n) (1 + t^2/n)^{(n+1)/2}} \quad (-\infty < t < \infty) \quad (9.2.1)$$

where the parameter n , as in the χ^2 -distribution, is a positive integer called the *number of degrees of freedom* of the distribution.

Theorem I. If X is a standard normal variate, χ^2 is χ^2 -distributed with n degrees of freedom, and X and χ^2 are independent, then

$Y = \frac{X}{\sqrt{\chi^2/n}} = \frac{X}{\sqrt{\chi^2}} \sqrt{n}$ has a *t-distribution* with n degrees of freedom.

Proof. We write $\frac{Y^2}{n} = \frac{\frac{1}{2}X^2}{\frac{1}{2}\chi^2}$.

Since X and χ^2 are independent, $\frac{1}{2}X^2$ and $\frac{1}{2}\chi^2$ are also independent, and both are γ -variates, the former with parameter $\frac{1}{2}$ and the latter with parameter $\frac{1}{2}n$, and hence their quotient Y^2/n is a $\beta_2(\frac{1}{2}, \frac{1}{2}n)$ variate (cf. Ex. 4 Sec. 6.6), so that the probability differential

$$\begin{aligned} dF &= \frac{(y^2/n)^{-1/2}}{B(\frac{1}{2}, \frac{1}{2}n)(1+y^2/n)^{(n+1)/2}} d(y^2/n) \\ &= \frac{2}{\sqrt{n}B(\frac{1}{2}, \frac{1}{2}n)(1+y^2/n)^{(n+1)/2}} dy \end{aligned}$$

Now as y ranges from $-\infty$ to ∞ , y^2/n traverses the interval $(0, \infty)$ twice, and hence

$$f_y(y) = \frac{1}{\sqrt{n}B(\frac{1}{2}, \frac{1}{2}n)(1+y^2/n)^{(n+1)/2}} \quad (-\infty < y < \infty)$$

Hence the theorem.

Characteristics of the t -distribution

The t -distribution is symmetrical about the origin. For $n=1$, $f(t) = 1/\pi(1+t^2)$ which is, in fact, the density function of the Cauchy distribution with parameters $\lambda=1$ and $\mu=0$; for this distribution,

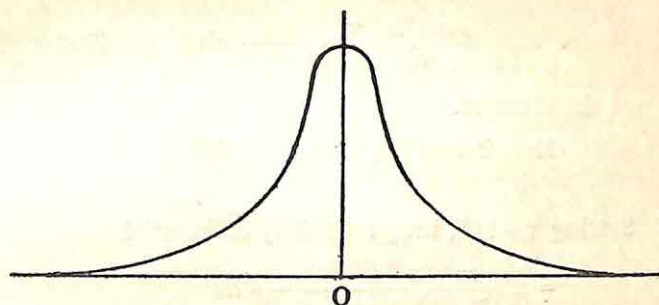


Fig. 21. t -Density Curve

we know, the mean does not exist. For $n > 1$, however, the mean of the t -distribution exists and, on account of symmetry, the mean, mode and the median are all zero.

9.3 *F*-DISTRIBUTION

Here

$$f(F) = \frac{m^{m/2} n^{n/2} F^{m/2-1}}{B(\frac{1}{2}m, \frac{1}{2}n)(mF+n)^{(m+n)/2}} \quad F > 0$$

$$= 0 \quad F < 0 \quad (9.3.1)$$

where m, n , both positive integers, are the two parameters of the distribution. The random variable will be called an $F(m, n)$ variate. We note that the distribution is not symmetrical with respect to the parameters m and n .

Theorem I. If χ_1^2 and χ_2^2 are independent variates having χ^2 -distribution with m and n degrees of freedom respectively, then

$$X = \frac{\chi_1^2/m}{\chi_2^2/n} = \frac{n\chi_1^2}{m\chi_2^2}$$

is an $F(m, n)$ variate.

Proof. Write $\frac{m}{n}X = \frac{\frac{1}{2}\chi_1^2}{\frac{1}{2}\chi_2^2}$.

Now $\frac{1}{2}\chi_1^2$ and $\frac{1}{2}\chi_2^2$ are independent $\gamma(\frac{1}{2}m)$ and $\gamma(\frac{1}{2}n)$ variates respectively, and hence $\frac{m}{n}X$ has a $\beta_2(\frac{1}{2}m, \frac{1}{2}n)$ distribution. Hence

$$dF = \frac{(mx/n)^{n/2-1}}{B(\frac{1}{2}m, \frac{1}{2}n)(1+mx/n)^{(m+n)/2}} d(mx/n)$$

$$= \frac{m^{n/2}n^{n/2}x^{n/2-1}}{B(\frac{1}{2}m, \frac{1}{2}n)(mx+n)^{(m+n)/2}} dx \quad (0 < x < \infty)$$

which proves the theorem.

Theorem II. If F is an $F(m, n)$ variate, then $X = 1/F$ is an $F(n, m)$ variate.

Proof Setting $x = 1/F$, the probability differential

$$= \frac{m^{m/2}n^{n/2}F^{m/2-1}}{B(\frac{1}{2}m, \frac{1}{2}n)(mF+n)^{(m+n)/2}} dF$$

$$= \frac{m^{m/2}n^{n/2}F^{m/2-1}}{B(\frac{1}{2}m, \frac{1}{2}n)(mF+n)^{(m+n)/2}} \left| \frac{dF}{dx} \right| dx$$

$$= \frac{m^{n/2}n^{n/2}x^{n/2-1}}{B(\frac{1}{2}n, \frac{1}{2}m)(nx+m)^{(m+n)/2}} dx \quad (0 < x < \infty)$$

Hence the theorem.

Characteristics of the F -distribution

It may be easily seen that the mean exists only for $n > 2$, and its value is $\frac{n}{n-2}$ which is independent of the parameter m and is greater than 1.

For $m > 2$, the distribution has a unique *mode* at the point $\frac{n(m-2)}{m(n+2)} < 1$.

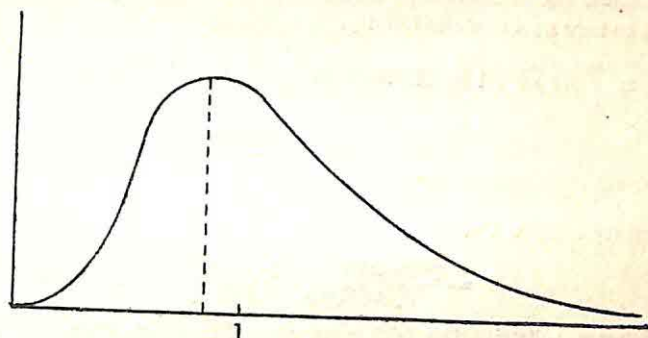


Fig. 22. Typical F -Density Curve

9.4 EXERCISES

1. If X_1, X_2, \dots, X_n are mutually independent normal variates each having mean zero and standard deviation σ , find the distribution of the sum of their squares.

2. Assume that the three velocity components (V_x, V_y, V_z) of any molecule of a gas are mutually independent random variables, each being normal $(0, \sqrt{kT/m})$ where k is Boltzmann's constant, m the mass of a molecule and T the absolute temperature of the gas. Prove that magnitude of the velocity V has the Maxwell-Boltzmann probability density function

$$av^2 e^{-\beta v^2} \quad (0 < v < \infty)$$

where

$$a = \sqrt{\frac{2}{\pi}} \left(\frac{m}{kT} \right)^{3/2}, \quad \beta = \frac{m}{2kT}$$

Find also the distribution of the kinetic energy of a gas molecule.

3. If (X, Y) has the general bivariate normal distribution, show that

$$\left\{ \frac{(X-m_x)^2}{\sigma_x^2} - 2\rho \frac{(X-m_x)(Y-m_y)}{\sigma_x \sigma_y} + \frac{(Y-m_y)^2}{\sigma_y^2} \right\} / (1-\rho^2)$$

has a χ^2 -distribution with 2 degrees of freedom.

4. Show, using Theorem I Sec. 9.1, that for the χ^2 -distribution with n degrees of freedom $\kappa_k = 2^{k-1}(k-1)!/n$, and hence obtain the coefficient of skewness γ_1 and the coefficient of excess γ_2 .

5. If X and Y are independent variates, X being χ^2 -distributed with m degrees of freedom and their sum $X+Y$ χ^2 -distributed with $m+n$ degrees of freedom, then show that Y is χ^2 -distributed with n degrees of freedom.

6. For the t -distribution with n degrees of freedom, prove that the variance exists only for $n > 2$ and that its value is $n/(n-2)$.

7. Calculate the coefficient of excess γ_2 for the $t(n)$ distribution, and show that it tends to zero as n tends to infinity.

8. From $\int_0^\infty f(F) dF = 1$ obtain the identity

$$\int_0^\infty \frac{F^{m/2-1} dF}{(mF+n)^{(m+n)/2}} = \frac{B(\frac{1}{2}m, \frac{1}{2}n)}{m^{m/2} n^{n/2}}$$

Use this identity to show that

$$a_k = \frac{\Gamma(\frac{1}{2}m+k)\Gamma(\frac{1}{2}n-k)}{\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} \left(\frac{n}{m}\right)^k \quad (k < \frac{1}{2}n)$$

9. If t has a t -distribution with n degrees of freedom, then show that t^2 is an $F(1, n)$ variate.

10. If χ_1^2, χ_2^2 are independent χ^2 -variates having m and n degrees of freedom respectively, find the distribution of χ_1^2/χ_2^2 .

CONVERGENCE 'IN PROBABILITY'

We start with the following fundamental inequality.

10.1 TCHEBYCHEFF'S INEQUALITY

If X is any random variable having a finite variance, then for any $\varepsilon > 0$

$$P(|X - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \quad (10.1.1)$$

where m and σ respectively denote the mean and standard deviation of X . (We note that the existence of the variance implies that of the mean.)

Proof. CONTINUOUS CASE. We have

$$P(|X - m| \geq \varepsilon) = \int_{|x - m| \geq \varepsilon} f(x) dx$$

Now in the range of integration $1 \leq (x - m)^2 / \varepsilon^2$, and hence

$$\text{R.H.S.} \leq \frac{1}{\varepsilon^2} \int_{|x - m| \geq \varepsilon} (x - m)^2 f(x) dx$$

Since the integrand is nonnegative

$$\text{R.H.S.} \leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} (x - m)^2 f(x) dx = \frac{\sigma^2}{\varepsilon^2}$$

DISCRETE CASE. The proof is similar to that in the continuous case.

$$\begin{aligned} P(|X - m| \geq \varepsilon) &= \sum_{|x_i - m| \geq \varepsilon} f_i \\ &\leq \frac{1}{\varepsilon^2} \sum_{|x_i - m| \geq \varepsilon} (x_i - m)^2 f_i \\ &\leq \frac{1}{\varepsilon^2} \sum_{i=-\infty}^{\infty} (x_i - m)^2 f_i = \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

The inequality (10.1.1) may also be written as

$$P(|X - m| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2} \quad (10.1.2)$$

for any $\varepsilon > 0$, or

$$P(|X - m| \geq \tau\sigma) \leq \frac{1}{\tau^2} \quad (10.1.3)$$

for any $\tau > 0$.

Remark. Tchebycheff's inequality brings out the significance of the variance as a measure of dispersion about the mean somewhat quantitatively. It states that the amount of probability mass outside the interval $(m - \varepsilon, m + \varepsilon)$ is less than or equal to σ^2/ε^2 which is obviously small, for a given ε , if the variance is small.

Example. Let X be normal (m, σ) . Then by Tchebycheff's inequality (10.1.3)

$$P(|X - m| \geq 2\sigma) \leq \frac{1}{4}$$

But, since $X^* = (X - m)/\sigma$ is normal $(0, 1)$, we have

$$P(|X - m| \geq 2\sigma) = P(|X^*| \geq 2) = .0456$$

the numerical value being obtained from Table I at the end of the book. This, however, shows that the Tchebycheff's inequality gives a rather poor bound for the probability in question.

10.2 CONVERGENCE 'IN PROBABILITY'

We shall now introduce a new concept of convergence, viz. convergence in probability or stochastic convergence which is defined as follows.

A sequence of random variables $X_1, X_2, \dots, X_n, \dots$ is said to converge in probability to a constant a , if for any $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - a| < \varepsilon) = 1 \quad (10.2.1)$$

or its equivalent

$$\lim_{n \rightarrow \infty} P(|X_n - a| \geq \varepsilon) = 0 \quad (10.2.2)$$

and we write

$$X_n \xrightarrow{\text{in } p} a \text{ as } n \rightarrow \infty$$

That is, speaking practically, as n increases the probability mass of the distribution of X_n accumulates more and more about the point a .

If there exists a random variable X such that $X_n - X \xrightarrow{\text{in } p} 0$ as $n \rightarrow \infty$, then we say that the given sequence of random variables *converges in probability to the random variable X* .

Remark. If a sequence of constants $a_n \rightarrow a$ as $n \rightarrow \infty$, then regarding a constant as a random variable having a one-point distribution at that point, we may also write $a_n \xrightarrow{\text{in } p} a$ as $n \rightarrow \infty$.

Although the concept of convergence in probability is basically different from that of ordinary convergence of a sequence of numbers, the following simple rules hold for convergence in probability as well.

Let $X_n \xrightarrow{\text{in } p} a$ and $Y_n \xrightarrow{\text{in } p} b$ as $n \rightarrow \infty$. Then as $n \rightarrow \infty$

$$(i) \quad X_n \pm Y_n \xrightarrow{\text{in } p} a \pm b \quad (10.2.3)$$

$$(ii) \quad X_n Y_n \xrightarrow{\text{in } p} ab \quad (10.2.4)$$

$$(iii) \quad X_n/Y_n \xrightarrow{\text{in } p} a/b, \quad \text{provided } b \neq 0 \quad (10.2.5)$$

Proof. Let A, B, C denote the events $|X_n - a| \geq \frac{1}{2}\epsilon$, $|Y_n - b| \geq \frac{1}{2}\epsilon$ and $|(X_n \pm Y_n) - (a \pm b)| \geq \epsilon$ respectively for any given $\epsilon > 0$. Then the complementary events $\bar{A}, \bar{B}, \bar{C}$ are respectively $|X_n - a| < \frac{1}{2}\epsilon$, $|Y_n - b| < \frac{1}{2}\epsilon$ and $|(X_n \pm Y_n) - (a \pm b)| < \epsilon$. If \bar{A} and \bar{B} occur simultaneously, then

$$|(X_n \pm Y_n) - (a \pm b)| \leq |X_n - a| + |Y_n - b| < \frac{1}{2}\epsilon + \frac{1}{2}\epsilon = \epsilon$$

i.e. \bar{C} occurs, so that $\bar{A}\bar{B}$ implies \bar{C} or $\bar{A}\bar{B} \subseteq \bar{C}$ or $C \subseteq A+B$. Hence

$$P(C) \leq P(A+B) \leq P(A) + P(B) \xrightarrow{n \rightarrow \infty} 0$$

since $X_n \xrightarrow{\text{in } p} a$, $Y_n \xrightarrow{\text{in } p} b$ as $n \rightarrow \infty$, so that $P(C) \rightarrow 0$ as $n \rightarrow \infty$. This shows that $X_n \pm Y_n \xrightarrow{\text{in } p} a \pm b$ as $n \rightarrow \infty$.

By induction the above rule may be extended to a finite number of sequences of random variables.

Next we prove that $cX_n \xrightarrow{\text{in } p} ca$ as $n \rightarrow \infty$, c being a constant. If $c = 0$, this is obvious. If $c \neq 0$, for any $\epsilon > 0$

$P(|cX_n - ca| \geq \epsilon) = P(|X_n - a| \geq \epsilon/|c|) \rightarrow 0$ as $n \rightarrow \infty$
 as $X_n \xrightarrow{\text{in } p} a$ as $n \rightarrow \infty$, which proves the proposition.

If $Z_n \xrightarrow{\text{in } p} 0$ as $n \rightarrow \infty$, then $Z_n^2 \xrightarrow{\text{in } p} 0$ as $n \rightarrow \infty$, for

$$P(Z_n^2 \geq \epsilon) = P(|Z_n| \geq \sqrt{\epsilon}) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Then if, as $n \rightarrow \infty$, $X_n \xrightarrow{\text{in } p} a$, then $X_n^2 \xrightarrow{\text{in } p} a^2$, for

$$X_n^2 = (X_n - a)^2 + 2a(X_n - a) + a^2 \xrightarrow{\text{in } p} a^2 \quad \text{as } n \rightarrow \infty$$

To prove (10.2.4) we note that

$$\begin{aligned} X_n Y_n &= \frac{1}{4}[(X_n + Y_n)^2 - (X_n - Y_n)^2] \\ &\xrightarrow{\text{in } p} \frac{1}{4}[(a+b)^2 - (a-b)^2] = ab \quad \text{as } n \rightarrow \infty \end{aligned}$$

(10.2.5) will follow from (10.2.4) if we can show that $\frac{1}{Y_n} \xrightarrow{\text{in } p} \frac{1}{b}$ as $n \rightarrow \infty$ ($b \neq 0$). To prove this we have

$$\left| \frac{1}{Y_n} - \frac{1}{b} \right| \leq \frac{|Y_n - b|}{|b|(|b| - |Y_n - b|)}$$

if $|Y_n - b| < |b|$. Let A denote the event $|Y_n - b| < |b|$ so that \bar{A} is the event $|Y_n - b| \geq |b|$, and let, for a given $\epsilon > 0$, B denote the event

$\left| \frac{1}{Y_n} - \frac{1}{b} \right| \geq \epsilon$. If AB occurs, i.e. A and B occur simultaneously, then

$$\frac{|Y_n - b|}{|b|(|b| - |Y_n - b|)} \geq \epsilon$$

or

$$|Y_n - b| \geq \frac{\epsilon b^2}{1 + \epsilon |b|}$$

Representing the last event by C , it follows that $AB \subseteq C$, and

$$B = AB + \bar{A}B \subseteq C + \bar{A}$$

So

$$P(B) \leq P(C + \bar{A}) \leq P(C) + P(\bar{A}) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

since $Y_n \xrightarrow{\text{in } p} b$ as $n \rightarrow \infty$, so that $P(B) \rightarrow 0$ as $n \rightarrow \infty$. This completes the proof.

As an immediate consequence of Tchebycheff's inequality, we have the following theorem regarding convergence in probability.

Tchebycheff's theorem. Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of random variables such that the mean m_n and standard deviation σ_n of X_n exist for all n . If $\sigma_n \rightarrow 0$ as $n \rightarrow \infty$, then

$$X_n - m_n \xrightarrow{\text{in } p} 0 \quad \text{as } n \rightarrow \infty$$

Proof. By (10.1.1), for any $\varepsilon > 0$,

$$P(|X_n - m_n| \geq \varepsilon) \leq \sigma_n^2 / \varepsilon^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

if $\sigma_n \rightarrow 0$ as $n \rightarrow \infty$. Hence $X_n - m_n \xrightarrow{\text{in } p} 0$ as $n \rightarrow \infty$.

COROLLARY. If, moreover, $m_n \rightarrow m$ as $n \rightarrow \infty$, then by the rule (10.2.3) together with the preceding remark $X_n \xrightarrow{\text{in } p} m$ as $n \rightarrow \infty$.

Bernoulli's theorem. If X_n is a binomial (n, p) variate, then

$$\frac{X_n}{n} \xrightarrow{\text{in } p} p \quad \text{as } n \rightarrow \infty \quad (10.2.6)$$

Proof. Since X_n is binomial (n, p)

$$E(X_n) = np, \quad \sigma(X_n) = \sqrt{npq} \quad (q = 1 - p)$$

Now for the sequence X_n/n , we have

$$E(X_n/n) = p, \quad \sigma(X_n/n) = \sqrt{pq/n} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Hence, by Tchebycheff's theorem, $X_n/n - p \xrightarrow{\text{in } p} 0$ as $n \rightarrow \infty$ whence the theorem follows.

If X_n denotes the number of successes in a Bernoullian sequence of n trials with probability of success p , then X_n is binomial (n, p) , and $f = X_n/n$ is the frequency ratio of successes, in terms of which we get another version of the above theorem stated as follows.

If f is the frequency ratio of successes in a Bernoullian sequence of n trials with probability of success p , then

$$f \xrightarrow{\text{in } p} p \quad \text{as } n \rightarrow \infty \quad (10.2.7)$$

Consider now any random experiment E in general. We stated in the frequency interpretation of probability that if E is repeated under uniform conditions a large number of times, the frequency ratio

of any event will be approximately equal to its probability. This practical sort of statement may be given a precise mathematical form by means of Bernoulli's theorem. Let the random variable $n(A)$ denote the frequency of any event A in a sequence of n repetitions of E so that its frequency ratio $f(A) = n(A)/n$. If we now mathematically interpret a sequence of repetitions of E under uniform conditions as a sequence of independent trials of E , it follows that $n(A)$ is binomially distributed with parameters $\{n, P(A)\}$, and Bernoulli's theorem states

$$f(A) \xrightarrow[\text{in } p]{} P(A) \quad \text{as } n \rightarrow \infty \quad (10.2.8)$$

Remark. It will be interesting to compare Bernoulli's theorem (10.2.8) with the frequency definition of probability (2.3.2). As the number of repetitions of E increases, the latter states that we are certain that the frequency ratio of an event gets closer and closer to its probability, whereas the former states that we become more and more sure that the frequency ratio will lie in a fixed small neighbourhood of the probability.

Law of large numbers. Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of random variables such that $S_n = X_1 + X_2 + \dots + X_n$ has a finite mean M_n and standard deviation Σ_n for all n . If $\Sigma_n = o(n)$, i.e. $\Sigma_n/n \rightarrow 0$ as $n \rightarrow \infty$, then

$$\frac{S_n - M_n}{n} \xrightarrow[\text{in } p]{} 0 \quad \text{as } n \rightarrow \infty \quad (10.2.9)$$

Proof. We have

$$E\left(\frac{S_n - M_n}{n}\right) = 0, \quad \sigma\left(\frac{S_n - M_n}{n}\right) = \frac{\Sigma_n}{n}$$

Hence if $\Sigma_n/n \rightarrow 0$ as $n \rightarrow \infty$, (10.2.9) follows from Tchebycheff's theorem.

If m_1, m_2, \dots respectively denote the means of X_1, X_2, \dots , then $M_n = m_1 + m_2 + \dots + m_n$, and writing

$$\bar{X} = (X_1 + X_2 + \dots + X_n)/n = S_n/n$$

$$\bar{m} = (m_1 + m_2 + \dots + m_n)/n = M_n/n$$

(10.2.9) may be written in the alternative form :

$$\bar{X} - \bar{m} \xrightarrow[\text{in } p]{} 0 \quad \text{as } n \rightarrow \infty \quad (10.2.10)$$

CASE OF EQUAL COMPONENTS. If the random variables X_1, X_2, \dots all have the same distribution with existent mean m and standard deviation σ , and X_1, X_2, \dots, X_n are mutually independent for all n , then $M_n = nm$ and $\Sigma_n = \sqrt{n}\sigma = o(n)$. Hence the law of large numbers holds, and, moreover, the form (10.2.10) simplifies to

$$\bar{X} \xrightarrow[\text{in } p]{} m \quad \text{as } n \rightarrow \infty \quad (10.2.11)$$

Remark. We note that the condition $\Sigma_n = o(n)$ is sufficient for holding of the law of large numbers but is not necessary. For the case equal of components, the condition of finiteness of the standard deviation σ (which is, however, essential for the above method of proof) is also not necessary. It can be shown that the law of large numbers for equal components holds under the simple condition that the mean m of the common distribution exists.

10.3 EXERCISES

1. Prove the following generalisation of Tchebycheff's inequality :

If X possesses a finite second order moment and c is any fixed number, then, for any $\epsilon > 0$, $P(|X - c| \geq \epsilon) \leq E\{(X - c)^2\}/\epsilon^2$.

2. If X is a nonnegative random variable having mean m , prove that, for any $\tau > 0$, $P(X \geq \tau m) \leq 1/\tau$.

3. Show, by Tchebycheff's inequality, that in 2,000 throws with a coin the probability that the number of heads lies between 900 and 1,100 is at least 19/20.

4. A random variable X has probability density function $12x^2(1-x)$ ($0 < x < 1$). Compute $P(|X - m| \geq 2\sigma)$, and compare it with the limit given by Tchebycheff's inequality.

5. If X is a $\gamma(n)$ variate, then show that

$$P(0 < X < 2n) \leq (n-1)/n$$

6. Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of pairwise uncorrelated random variables having finite means $m_1, m_2, \dots, m_n, \dots$ and standard deviations $\sigma_1, \sigma_2, \dots, \sigma_n, \dots$ respectively. Show that the law of large numbers holds if the sequence $\{\sigma_n\}$ is bounded.

7. Obtain Bernoulli's theorem as a particular case of the law of large numbers for equal components.

8. In a Poisson sequence of n trials, if f denotes the frequency ratio of successes and $\bar{p} = \frac{1}{n} \sum p_i$, where p_i is the probability of success in the i th trial ($i = 1, 2, \dots, n$), then prove that $f - \bar{p} \xrightarrow[\text{in } p]{} 0$ as $n \rightarrow \infty$.

LIMIT THEOREMS

11.1 NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

We have already considered one approximation to the binomial (n, p) distribution, viz. the Poisson approximation which holds if p is small and n large such that np is of moderate magnitude. But if p is not small but of moderate magnitude and n is large, then we shall show that the binomial distribution approximates to the continuous normal distribution with parameters (np, \sqrt{npq}) . This is a very important result which is precisely stated in the following theorem.

DeMoivre-Laplace limit theorem. Let X_n be a binomial (n, p) variate ($0 < p < 1$), the corresponding standardised variate being

$$X_n^* = \frac{X_n - np}{\sqrt{npq}} \quad (q = 1 - p)$$

Then for any fixed numbers $a, b (> a)$

$$\lim_{n \rightarrow \infty} P(a < X_n^* \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \quad (11.1.1)$$

Proof. The spectrum of X_n consists of the points $0, 1, 2, \dots, n$, and

$$f_i = P(X_n = i) = \binom{n}{i} p^i q^{n-i} \quad (i = 0, 1, \dots, n) \quad (i)$$

Hence the spectrum points x_i^* of X^* are given by

$$x_i^* = \frac{i - np}{\sqrt{npq}} \quad (i = 0, 1, \dots, n) \quad (ii)$$

and

$$P(X_n^* = x_i^*) = P(X_n = i) = f_i \quad (iii)$$

Then

$$\Delta x_i^* = x_{i+1}^* - x_i^* = \frac{1}{\sqrt{npq}} \quad (iv)$$

We note that $\Delta x_i^* \rightarrow 0$ as $n \rightarrow \infty$, and this explains how the discrete spectrum of X_n^* tends to a continuous spectrum over the entire real axis as $n \rightarrow \infty$. We have

$$P(a < X_n^* \leq b) = \sum_{a < x_i^* \leq b} f_i \quad (v)$$

By (ii)

$$i = np + x_i^* \sqrt{npq}, \quad n-i = nq - x_i^* \sqrt{npq}$$

For $a < x_i^* \leq b$, a, b being fixed numbers, i and $n-i$ both $\rightarrow \infty$ as $n \rightarrow \infty$. Then

$\log f_i = \log n! - \log i! - \log (n-i)! + i \log p + (n-i) \log q$
and using Stirling's formula :

$$\log v! = \log \sqrt{2\pi} + (v + \frac{1}{2}) \log v - v + \theta/12v \quad (0 < \theta < 1)$$

we get

$$\begin{aligned} \log f_i &= -\log \sqrt{2\pi npq} - (i + \frac{1}{2}) \left(\frac{i}{np} \right) - (n-i + \frac{1}{2}) \log \left(\frac{n-i}{nq} \right) \\ &\quad + \frac{1}{12} \left(\frac{\theta_1}{n} - \frac{\theta_2}{i} - \frac{\theta_3}{n-i} \right) \quad (0 < \theta_1, \theta_2, \theta_3 < 1) \end{aligned} \quad (vi)$$

Now

$$\frac{i}{np} = 1 + x_i^* \sqrt{\frac{q}{np}}, \quad \frac{n-i}{nq} = 1 - x_i^* \sqrt{\frac{p}{nq}}$$

We know, from Taylor's theorem, that for $|x| < \frac{1}{2}$

$$\log(1+x) = x - \frac{1}{2}x^2 + \lambda x^3 \quad (|\lambda| < 1)$$

For sufficiently large n , $x_i^* \sqrt{q/np}$ and $x_i^* \sqrt{p/nq}$ can be made numerically less than $\frac{1}{4}$, and hence

$$\begin{aligned} \log \left(\frac{i}{np} \right) &= \log \left(1 + x_i^* \sqrt{\frac{q}{np}} \right) \\ &= x_i^* \sqrt{\frac{q}{np}} - \frac{x_i^{*2}}{2} \frac{q}{np} + \lambda_1 x_i^{*3} \left(\frac{q}{np} \right)^{3/2} \end{aligned}$$

and

$$\begin{aligned}\log \left(\frac{n-i}{nq} \right) &= \log \left(1 - x_i^* \sqrt{\frac{p}{nq}} \right) \\ &= -x_i^* \sqrt{\frac{p}{nq}} - \frac{x_i^{*2}}{2} \frac{p}{nq} + \lambda_2 x_i^{*3} \left(\frac{p}{nq} \right)^{3/2} \\ &\quad (|\lambda_1|, |\lambda_2| < 1)\end{aligned}$$

Inserting these in (vi), we can write

$$\log f_i = -\log \sqrt{2\pi npq} - \frac{x_i^{*2}}{2} + \frac{r}{\sqrt{n}}$$

where $r = r(n, x_i^*)$ is such that $|r| < A$, A being a constant independent of n and x_i^* . Hence

$$f_i = \frac{1}{\sqrt{2\pi npq}} e^{-x_i^{*2}/2} + \frac{R}{n}$$

where $R = R(n, x_i^*)$ is such that $|R| < B$, a constant.

Now by (iv) and (v)

$$P(a < X_n^* \leq b) = \frac{1}{\sqrt{2\pi}} \sum_{a < x_i^* \leq b} e^{-x_i^{*2}/2} \Delta x_i^* + \frac{1}{n} \sum_{a < x_i^* \leq b} R$$

Since $\Delta x_i^* = 1/\sqrt{npq}$, the number of terms in the summation is $\leq (b-a)\sqrt{npq} + 1$, and hence

$$\left| \frac{1}{n} \sum_{a < x_i^* \leq b} R \right| < \frac{B}{n} \{ (b-a)\sqrt{npq} + 1 \} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Therefore, the second sum $\rightarrow 0$ as $n \rightarrow \infty$, so that

$$\begin{aligned}\lim_{n \rightarrow \infty} P(a < X_n^* \leq b) &= \frac{1}{\sqrt{2\pi}} \lim_{n \rightarrow \infty} \sum_{a < x_i^* \leq b} e^{-x_i^{*2}/2} \\ &= \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx\end{aligned}$$

This completes the proof of the theorem.

As introduced in (5.8.5), let $\phi(x)$ and $\varphi(x)$ denote the standard normal density and distribution functions respectively, in terms of which (11.1.1) may be written as

$$\lim_{n \rightarrow \infty} P(a < X_n^* \leq b) = \int_a^b \phi(x) dx = \varphi(b) - \varphi(a) \quad (11.1.2)$$

If $F_n(x)$ denotes the distribution function of X_n and $F_n^*(x)$ that of X_n^* , then making $a \rightarrow -\infty$ and replacing b by x we get

$$\lim_{n \rightarrow \infty} F_n^*(x) = \varphi(x) \quad (11.1.3)$$

Also (11.1.3) clearly implies (11.1.2), i.e. (11.1.3) gives an equivalent form of the above theorem, which states that the distribution function of the standardised binomial (n, p) distribution tends to the standard normal distribution function as n tends to infinity, provided p is kept fixed.

In terms of the variate X_n (11.1.2) may be written in the form

$$\lim_{n \rightarrow \infty} P(np + a\sqrt{npq} < X_n \leq np + b\sqrt{npq}) = \int_a^b \phi(x) dx \quad (11.1.4)$$

A working statement of (11.1.4) will be that if n is large and p is of moderate magnitude, we have the approximation formula

$$P(np + a\sqrt{npq} < X_n \leq np + b\sqrt{npq}) \simeq \int_a^b \phi(x) dx \quad (11.1.5)$$

Example. If a die is thrown 1,800 times, find the probability that the frequency of the event 'multiple of three' lies between 600 ± 50 .

The frequency of the given event is binomially distributed with parameters $n=1,800$ and $p=\frac{1}{3}$. Here p is not small and n is large so that we can use the normal approximation. We have $np=600$, $\sqrt{npq}=20$, and, putting $a=-2.5$ and $b=2.5$ in (11.1.5), the required probability approximately equals

$$\int_{-2.5}^{2.5} \phi(x) dx = 0.988$$

(The numerical value of the integral is obtained from Table I at the end of the book.)

For a graphical representation of this approximation, we note that the probability differential $\phi(x)dx$ corresponds to f_i so that $\phi(x)$ will correspond to $f_i/\Delta x_i^* = f_i/\sqrt{npq}$. Fig. 23 shows the modified probability diagram of the binomial distribution, in which the ordinates are f_i/\sqrt{npq} instead of f_i for $n=16$, $p=\frac{1}{2}$, together with the standard normal density curve.

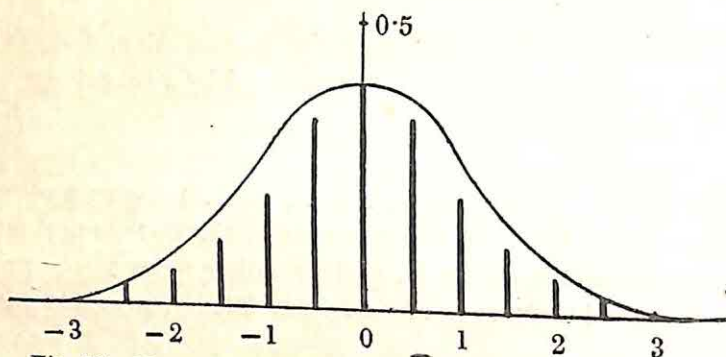


Fig. 23. Normal Approximation to the Binomial Distribution

11.2 FUNDAMENTAL LIMIT THEOREMS

For convenience of expression, we shall make use of the following terminology.

Asymptotically normal. Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of random variables, and $a_1, a_2, \dots, a_n, \dots$ and $\beta_1, \beta_2, \dots, \beta_n, \dots$ two sequences of constants. If the distribution function of the random variable $\frac{X_n - a_n}{\beta_n}$ tends to $\phi(x)$, the standard normal distribution function for all x , then X_n is said to be *asymptotically normal* (a_n, β_n) . It is, however, not implied that a_n, β_n respectively denote the mean and standard deviation of X_n ; these may not even exist. The practical sense of this expression is obviously that the distribution of X_n is approximately normal (a_n, β_n) for large values of n .

Now if X_1, X_2, \dots, X_n are mutually independent random variables each binomial $(1, p)$, then by the reproductive property of the binomial distribution, their sum $S_n = X_1 + X_2 + \dots + X_n$ is a binomial (n, p) variate, and DeMoivre-Laplace limit theorem states that S_n is asymptotically normal (np, \sqrt{npq}) . It is interesting to know that such a theorem holds not only for a sequence of binomial variates, but, in

general, for a large class of sequences of random variables obeying certain conditions. This is expressed by the following fundamental theorem known as the central limit theorem which we state without proof.

Central limit theorem. Let X_1, X_2, \dots, X_n be a sequence of random variables having finite variances such that X_1, X_2, \dots, X_n is a set of mutually independent random variables for all n , so that their sum

$$S_n = X_1 + X_2 + \dots + X_n$$

has a finite mean M_n and standard deviation Σ_n . Now a necessary and sufficient condition for S_n to be asymptotically normal (M_n, Σ_n) is that, for any given $\tau > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{\Sigma_n^2} \sum_{k=1}^n \int_{|x - m_k| \geq \tau \Sigma_n} (x - m_k)^2 f_k(x) dx = 0 \quad (11.2.1)$$

for the continuous case, or

$$\lim_{n \rightarrow \infty} \frac{1}{\Sigma_n^2} \sum_{k=1}^n \sum_{|x_i^{(k)} - m_k| \geq \tau \Sigma_n} \{x_i^{(k)} - m_k\}^2 f_i^{(k)} = 0 \quad (11.2.2)$$

for the discrete case where, m_k is the mean of X_k , $f_k(x)$ the density function of X_k for the continuous case, and for the discrete case $x_i^{(k)}$ denotes the general point of the spectrum of X_k having probability mass $f_i^{(k)}$. The above condition is called *Lindeberg's condition*.

This condition is, however, not very restrictive and is satisfied by many a sequence of random variables. In particular, we shall now show that the condition is satisfied for the case of equal components.

CASE OF EQUAL COMPONENTS. If the random variables X_1, X_2, \dots all have the same distribution with mean m and standard deviation σ , then Lindeberg's condition is fulfilled, and hence S_n is asymptotically normal ($nm, \sqrt{n}\sigma$), as in this case $M_n = nm$, $\Sigma_n = \sqrt{n}\sigma$.

Proof. Let us prove it for the continuous case, the proof for the discrete case being similar.

$$\text{L.H.S. of (11.2.1)} = \frac{1}{\sigma^2} \lim_{n \rightarrow \infty} \int_{|x - m| \geq \tau \sqrt{n}\sigma} (x - m)^2 f(x) dx$$

where $f(x)$ is the common density function of the random variables.

Since $\sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 f(x) dx$ exists, the R.H.S. is clearly zero for any fixed $\tau > 0$. Hence the proof.

We note $\frac{S_n - nm}{\sqrt{n}\sigma} = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$, and hence the central limit theorem for equal components may also be stated as : \bar{X} is asymptotically normal $(m, \sigma/\sqrt{n})$.

Remarks

1. The central limit theorem holds for equal components if simply their common distribution has a finite variance.

2. EXAMPLE OF FAILURE : CAUCHY DISTRIBUTION. If each random variable of the sequence X_1, X_2, \dots has a Cauchy distribution having parameters (λ, μ) , then we know that \bar{X} also has a Cauchy distribution having the same parameters (cf. Ex. 28 Sec. 6.8) and thus is not asymptotically normal. The violation of the central limit theorem is, however, a consequence of the fact that the Cauchy distribution has no finite mean and variance.

3. For the case of equal components, we can show that the central limit theorem implies the law of large numbers. If the sequence X_1, X_2, \dots obeys the central limit theorem, the distribution function of $\sqrt{n}(\bar{X} - m)/\sigma$ tends to $\phi(x)$ for all x . Let $a (> 0)$ be a given number. Then for any $\varepsilon > 0$,

$$\begin{aligned} P(|\bar{X} - m| \geq \varepsilon) &= P\left\{ \left| \frac{\sqrt{n}(\bar{X} - m)}{\sigma} \right| \geq \frac{\sqrt{n}\varepsilon}{\sigma} \right\} \\ &\leq P\left\{ \left| \frac{\sqrt{n}(\bar{X} - m)}{\sigma} \right| \geq a \right\} \quad \text{for } n > a^2 \sigma^2 / \varepsilon^2 \\ &\rightarrow 2\{1 - \phi(a)\} \quad \text{as } n \rightarrow \infty \end{aligned}$$

or

$$0 \leq \lim_{n \rightarrow \infty} P(|\bar{X} - m| \geq \varepsilon) \leq 2\{1 - \phi(a)\}$$

Since this holds for all a , we have making $a \rightarrow \infty$

$$\lim_{n \rightarrow \infty} P(|\bar{X} - m| \geq \varepsilon) = 0$$

which shows that $\bar{X} \xrightarrow{\text{in } p} m$ as $n \rightarrow \infty$, i.e. the law of large numbers holds.

The converse of this is not true, for the law of large number holds even for sequences of random variables for which the variance of their common distribution does not exist, and as such the central limit theorem cannot hold. These remarks do not, however, apply for the general case of non-identically distributed random variables.

We shall now state, without proof, another fundamental limit theorem concerning characteristic functions.

Limit theorem for characteristic functions. Let X_1, X_2, \dots, X_n ... be sequence of random variables having distribution functions $F_1(x), F_2(x), \dots, F_n(x)$... and characteristic functions $\chi_1(t), \chi_2(t), \dots, \chi_n(t)$... respectively. If as $n \rightarrow \infty$ $F_n(x) \rightarrow$ a distribution function $F(x)$ which determines the characteristic function $\chi(t)$, then $\chi_n(t) \rightarrow \chi(t)$ as $n \rightarrow \infty$. Conversely, if $\chi_n(t) \rightarrow$ a characteristic function $\chi(t)$ as $n \rightarrow \infty$, then $F_n(x) \rightarrow F(x)$, the distribution function corresponding to $\chi(t)$.

With the help of this theorem, we can prove with surprising ease the theorems on Poisson and normal limits of the binomial distribution.

POISSON DISTRIBUTION AS A LIMIT OF THE BINOMIAL DISTRIBUTION. The characteristic function $\chi_n(t)$ of the binomial (n, p) distribution is given by

$$\chi_n(t) = (pe^{it} + q)^n = \{1 + p(e^{it} - 1)\}^n$$

Set $p = \mu/n$, μ being a fixed positive number and make $n \rightarrow \infty$. Then

$$\chi_n(t) = \left\{1 + \frac{\mu}{n} (e^{it} - 1)\right\}^n \rightarrow e^{\mu(e^{it} - 1)}$$

which is the characteristic function of the Poisson- μ distribution. Hence it follows from the above theorem that the binomial $(n, \mu/n)$ ($\mu > 0$) distribution tends to the Poisson- μ distribution as $n \rightarrow \infty$.

DEMOIVRE-LAPLACE THEOREM. The characteristic function $\chi_n^*(t)$ of $X_n^* = (X_n - np) / \sqrt{npq}$ is given by

$$\begin{aligned} \chi_n^*(t) &= e^{-inpt/\sqrt{npq}} (p e^{it/\sqrt{npq}} + q)^n \\ &= (p e^{iat/\sqrt{npq}} + q e^{-ipt/\sqrt{npq}})^n \end{aligned}$$

We know, for any real x

$$e^{ix} = 1 + ix - \frac{x^2}{2!} + \theta \frac{x^3}{3!}$$

where θ is a complex quantity such that $|\theta| < 1$. Hence

$$e^{iqt/\sqrt{npq}} = 1 + \frac{iqt}{\sqrt{npq}} - \frac{q^2 t^2}{2npq} + \theta_1 \frac{q^3 t^3}{6(npq)^{3/2}}$$

and

$$e^{-ipt/\sqrt{npq}} = 1 - \frac{ipt}{\sqrt{npq}} - \frac{p^2 t^2}{2npq} + \theta_2 \frac{p^3 t^3}{6(npq)^{3/2}} \quad (|\theta_1|, |\theta_2| < 1)$$

so that

$$pe^{iqt/\sqrt{npq}} + qe^{-ipt/\sqrt{npq}} = 1 - \frac{t^2}{2n} + \theta \frac{t^3}{(npq)^{3/2}} \quad (|\theta| < 1)$$

and

$$\chi_n^*(t) = \left\{ 1 - \frac{t^2}{2n} + \theta \frac{t^3}{(npq)^{3/2}} \right\}^n \rightarrow e^{-t^2/2} \quad \text{as } n \rightarrow \infty$$

Since $e^{-t^2/2}$ is the characteristic function of the standard normal distribution, DeMoivre-Laplace limit theorem follows.

11.3 EXERCISES

1. Find, using the normal approximation, the probability that the number of heads in 2,000 throws with a coin lies between 900 and 1,100, and compare it with the lower limit given by Tchebycheff's inequality in Ex. 3 Sec. 10.3.

2. Find the number of times a die has to be thrown such that the probability that the difference between the frequency ratio of sixes and $1/6$ is in absolute value less than .01 is .99.

3. Show that the central limit theorem holds under the following sufficient condition called *Liapounoff's condition*: The random variables $X_1, X_2, \dots, X_n, \dots$ have finite third order moments such that

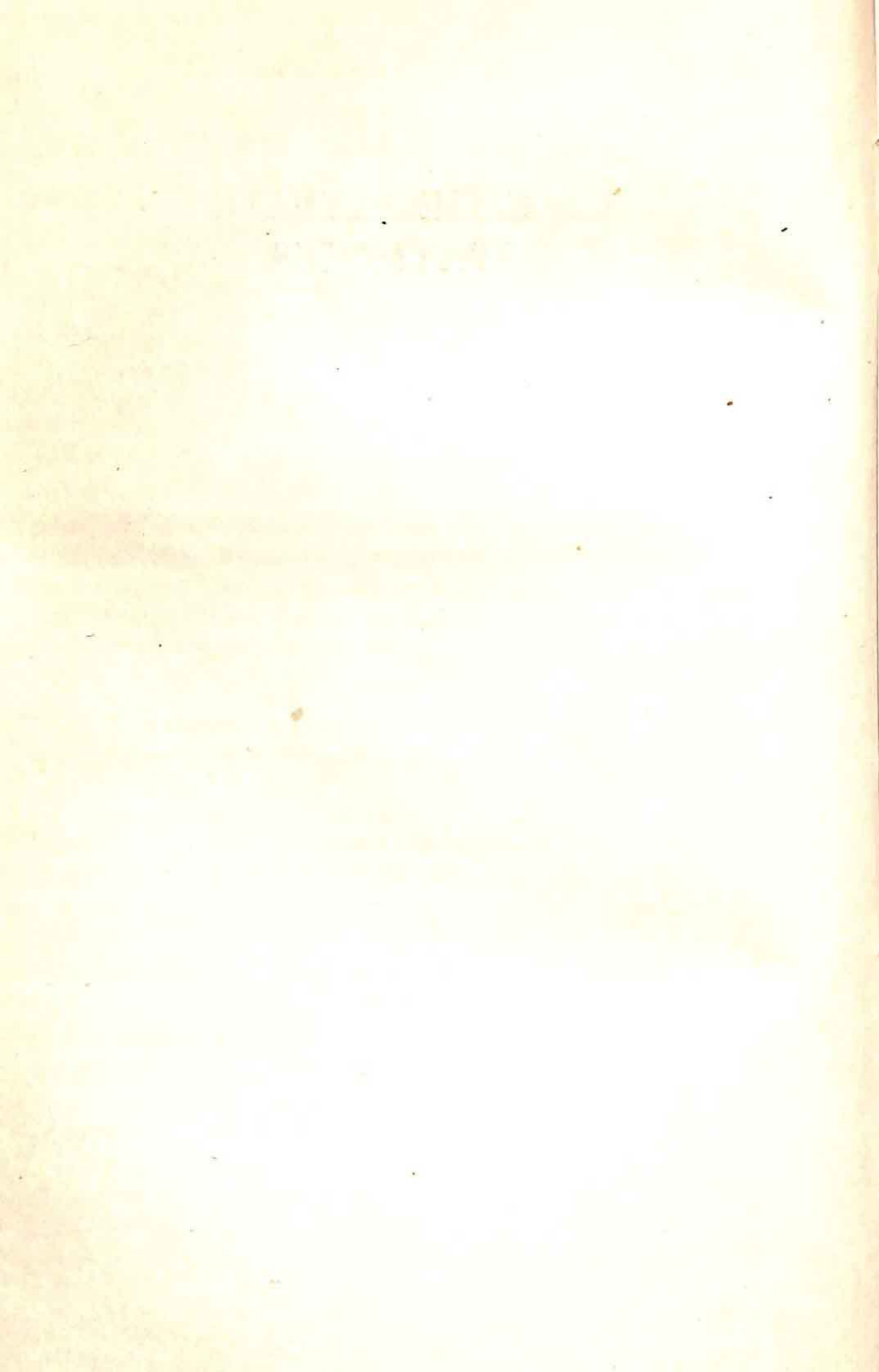
$$\lim_{n \rightarrow \infty} \frac{1}{\Sigma_{k=1}^n} \sum_{k=1}^n E(|X_k - m_k|^3) = 0$$

4. Prove that if Lindeberg's condition is fulfilled, $\Sigma_n \rightarrow \infty$ as $n \rightarrow \infty$.

5. Show that the density function of a standardised $\gamma(n)$ variate, n being a positive integer, tends to $\phi(x)$, the standard normal density function as $n \rightarrow \infty$, and hence deduce that a $\gamma(n)$ variate is asymptotically normal. (Assume, if necessary, the validity of interchanging limit as $n \rightarrow \infty$ and the sign of integration for proving the latter part.)

6. By the method of characteristic functions, show that (a) a χ^2 -variate with n degrees of freedom is asymptotically normal $(n, \sqrt{2n})$ and (b) a Poisson variate with parameter n (a positive integer) is asymptotically normal (n, \sqrt{n}) .

MATHEMATICAL
STATISTICS



RANDOM SAMPLES

12.1 POPULATIONS AND SAMPLES

Let E be a given random experiment and A any event connected with it. The basic problem of *statistics* may be regarded as the experimental determination of the probability of the event A . The method for this is readily obtained from the fundamental rule of interpretation of probabilities, viz. the *frequency interpretation* which states that if we repeat the experiment E under identical or uniform conditions a large number of times, the observed frequency ratio of A gives an approximate experimental value of the probability of A . Such a problem is, in general, called a problem of *estimation*. An estimate of the probability being thus obtained, we can employ it for predicting the rough values of future frequency ratios. Another type of problem may arise as follows. Suppose we know from a theoretical model or otherwise that the probability of A should have a given value, and we want to test experimentally if our conjecture is tenable or not. This problem which is closely allied to the problem of estimation is known as *testing of hypothesis*. Here also we are guided by the frequency interpretation; if we find that the frequency ratio of A for a long sequence of trials of E lies close to the suspected value of its probability, we may reasonably believe that our hypothesis is correct, and if it is otherwise, we should reject the hypothesis. Now, as we can obviously see, *exact* estimation is not possible, for in order to compute the exact value of the probability we must have to repeat the random experiment, so to say, an *infinite* number of times which is a practical impossibility. In practice, we have to depend on a finite number of repetitions of E which give only an approximate value of the probability. Naturally then, when an estimate is found, we shall also be interested in knowing how good is the estimate. Likewise, in the latter problem of testing of hypothesis exact decision is impossible ; we can only decide with varying degrees of *reasonableness*

or *confidence*, and it is necessary to supplement any such decision by a quantitative statement of the degree of this confidence.

The basic problems of statistics as stated in the above forms are rather trivial. But these acquire considerable complexity and importance when we go over to the probability distribution of a random variable X instead of simply the probability of an event. The probability distribution will be usually unknown, and it will be our problem to determine it experimentally or to test a hypothesis regarding the distribution. The mode of experimentation will, however, be same as before, viz. repeating the random experiment E under uniform conditions a large number of times and recording the result of each repetition. The frequency ratio of the event $X = \xi_k$, where ξ_k is any point of the spectrum of X in the discrete case, or that of the event $x < X \leq x + dx$ in the continuous case will then respectively approximate to $P(X = \xi_k) = f_k$ or the probability differential $f(x) dx$, and in this way we can roughly determine the distribution. Now if the random experiment is performed once, the result is evidently an event point, and corresponding to this event point we get a value of the random variable X , so that any particular performance of E gives a value of X , and hence a sequence of trials of E will give a sequence of values of X . The exact determination of the probability distribution will be provided, as we have remarked earlier, by a conceptual or hypothetical sequence of infinitely many repetitions of E under uniform conditions. This infinite sequence of repetitions of E will give rise to an infinite number of values of X , the totality of which will be called the *population* of the random variable X . A practical data is, however, obtained by repeating the experiment a finite number of times, say, n times which yields a sequence of n observed values :

$$x_1, x_2, \dots, x_n \quad (12.1.1)$$

of X , which is called a *sample of size n drawn from the population of X* . If now we perform another sequence of n trials of E , the sequence (12.1.1) will not be generally reproduced, but we shall get a different sequence of values of X : x_1', x_2', \dots, x_n' . Thus if different samples of size n are repeatedly drawn under uniform conditions, the sets of observed values of the random variable will fluctuate at random. In this sense, the sample is said to be a *random sample*.

The individual values x_1, x_2, \dots, x_n of the sample (12.1.1) are usually called the *sample values*. In statistical terminology, the distribution of the given random variable X is often referred to as the *distribution of the population*, e. g.. we shall speak of the distribution function of the population or the characteristics of the population to mean the corresponding quantities belonging to the random variable. The sample is thus the basic experimental material at our disposal, and the methods of extracting information about the population (i.e. about X) from the sample will form the subject-matter of statistical analysis.

Examples

1. Let E denote the random experiment of throwing a die and the random variable X the number on the die. If we imagine that the die is repeatedly thrown an infinitely large number of times, we would get an infinite sequence of observed values of X which forms the population of the random variable. If now the die is actually thrown 100 times, a sequence of 100 values of X is obtained, which is then a sample of size 100 from the population.

2. Let E consist in selecting at random a male person in the city of Calcutta in the age group of 20-25 years during a period of one week, and measuring his *height* which is represented by the random variable X . In order to repeat the experiment E under uniform conditions, it is obviously necessary to replace the selected individual in the given category before making the next selection, for otherwise the composition of the category (comparable to an urn containing individuals) and hence the conditions of the experiment E would alter visibly. The population of heights of persons belonging to the given category will be obtained by repeating the experiment E infinite number of times under uniform conditions. We remark that this ideal population of heights is conceptually different from the totality of heights of the different individuals in the category; the ideal population is always conceived as infinite, whereas the number of individuals in the given category is finite. If, however, the number of individuals is very large and may be taken to be infinite for practical purposes, as is presumably the case here, it follows from our discussions in Sec. 4.3 that it does not matter much if the successive drawings are made without replacements which is very often done in practice.

12.2 DISTRIBUTION OF THE SAMPLE

To describe stochastically the distribution of the sample values, we define a *fake* random variable $\overset{\circ}{X}$ which takes the sample values x_1, x_2, \dots, x_n each with probability $1/n$. (All the sample values are, however, not necessarily distinct; in case a particular sample value is repeated, say, 3 times in the sample, it gets a share of probability mass $3/n$.) This empirical probability distribution of $\overset{\circ}{X}$ which is always of the discrete type, is called the *distribution of the sample* or the *empirical distribution* as distinct from the theoretical distribution of the population. We shall now show the important fact that if the size of the sample is large, the distribution of the sample approximates to the distribution of the population.

Let $F(x)$ be the distribution function of the population and $\overset{\circ}{F}(x)$ that of the distribution of the sample. By definition

$$\overset{\circ}{F}(x) = P(\overset{\circ}{X} \leq x) = v/n$$

where v is the number of sample values $\leq x$. Now v/n precisely denotes the frequency ratio of the event $X \leq x$, and hence for large n

$$v/n \simeq P(X \leq x) = F(x)$$

or if n is large

$$\overset{\circ}{F}(x) \simeq F(x) \quad (12.2.1)$$

This is usually expressed by saying that the distribution of the sample is the *statistical image* of the distribution of the population.

12.3 TABLES AND GRAPHICAL REPRESENTATIONS

Cumulative graph. The distribution curve of $\overset{\circ}{X}$: $y = \overset{\circ}{F}(x)$, which is always a step curve, is called the *cumulative graph* or sometimes the *sum polygon* of the sample. For large samples, it follows from (12.2.1) that the cumulative graph will run close to the distribution curve of the population: $y = F(x)$.

Discrete population : Frequency diagram. When the sample is drawn from a seemingly discrete population, all the sample values x_1, x_2, \dots, x_n are usually not distinct. Then let the distinct sample values be denoted by $\xi_1, \xi_2, \dots, \xi_m$ and v_j be the frequency of ξ_j (i.e. v_j is

the number of times the value ξ_j is repeated in the sample) ($j = 1, 2, \dots, m$), so that

$$\sum v_j = n \quad (12.3.1)$$

The data (12.1.1) may now be conveniently presented in a tabular form in which ξ 's are entered in one column and the corresponding v 's in another, the sum of the v 's being equal to the size of the sample.

The spectrum of the empirical distribution then consists of the points $\xi_1, \xi_2, \dots, \xi_m$ and since v_j is the frequency of ξ_j , the probability mass at ξ_j is v_j/n . Hence the probability diagram of X is obtained by erecting an ordinate of height v_j/n at each point ξ_j . This is called the *frequency diagram* of the sample. We note that ξ 's are also points of the spectrum of the population (which may not, however, include all of them) and the frequency ratio of the event $X = \xi_j$ is v_j/n , and hence for large n

$$v_j/n \simeq P(X = \xi_j) = f_j \quad (12.3.2)$$

i.e. the ordinates of the frequency diagram are approximately equal to the corresponding ordinates of the probability diagram of the population for large samples.

Example 1. The random variable represents the number of counts of α -rays emitted by a radioactive source as recorded by a Geiger-Mueller counter for a given time interval, and the results for 3,455 intervals are given by the following table :

No. of counts	Frequency	No. of counts	Frequency
0	8	9	220
1	59	10	121
2	177	11	85
3	311	12	24
4	492	13	22
5	528	14	6
6	601	15	3
7	467		
8	331	Total	3455

Continuous population : Grouping of data. Histogram. If the population is continuous, all the sample values are more or less distinct, and a ξ - v table of the above type is practically useless. In such cases the data is usually reduced to a comprehensible form by a process known as *grouping into classes*. This is an important idea in practical statistics which may be described in outline as follows. Note the smallest and largest of the sample values, and select a *convenient* interval (a, b) containing all the sample values x_1, x_2, \dots, x_n . Divide (a, b) into $m(<n)$ *suitable* sub-intervals by the points t_0, t_1, \dots, t_m such that $a = t_0 < t_1 < t_2 < \dots < t_m = b$. The sub-intervals $(t_{j-1}, t_j]$ ($j = 1, 2, \dots, m$) are called the *class intervals* or simply *classes*, and the points t_j the *class limits*. The number of sample values belonging to each class interval $(t_{j-1}, t_j]$ is counted which will be called the corresponding *class frequency* and denoted by v_j . A table is then drawn up showing the class frequencies v_j against the class intervals $(t_{j-1}, t_j]$. A data presented in this form is called a *grouped data*.

Remarks

1. In a grouped table, however, we lose sight of the individual sample values and thereby lose somewhat in exactness of knowledge, but this is usually outweighed by the gain in advantage in handling the data.

2. The class intervals are often taken to be of equal length, but sometimes, depending on the nature of the data, unequal intervals are also used.

To obtain a graphical representation of a grouped data we proceed as follows. Let ξ_j denote the middle point of the j th class interval which is called the *class midpoint* or *class mark* and $\Delta\xi_j$ the length of the j th class, i.e.

$$\xi_j = \frac{1}{2}(t_{j-1} + t_j), \quad \Delta\xi_j = t_j - t_{j-1} \quad (j = 1, 2, \dots, m) \quad (12.3.3)$$

On the j th class interval $(t_{j-1}, t_j]$ draw a rectangle of height $v_j/n\Delta\xi_j$ for all j ; the resulting diagram will be called the *histogram* of the sample. The tops of the rectangles form a graph having equation

$$y = \overset{\circ}{f}(x) \quad (12.3.4)$$

where

$$\hat{f}(x) = \frac{1}{\Delta \hat{\xi}_j} \frac{v_j}{n} \quad \text{in } t_{j-1} < x \leq t_j \quad (j = 1, 2, \dots, m)$$

Now the area of the j th rectangle is v_j/n which is also the frequency ratio of the event $t_{j-1} < X \leq t_j$. Hence if the sample is sufficiently large and the class intervals sufficiently small, then

$$v_j/n \simeq P(t_{j-1} < X \leq t_j) = \int_{t_{j-1}}^{t_j} f(x) dx \simeq f(\hat{\xi}_j) \Delta \hat{\xi}_j$$

where $f(x)$ denotes the density function of the population.

Since again

$$\begin{aligned} v_j/n &= \hat{f}(\hat{\xi}_j) \Delta \hat{\xi}_j \\ \hat{f}(\hat{\xi}_j) &\simeq f(\hat{\xi}_j) \end{aligned} \quad (j = 1, 2, \dots, m)$$

or we may write

$$\hat{f}(x) \simeq f(x) \quad (12.3.5)$$

Thus the upper part of the histogram is the statistical image of the density curve of the population.

Example 2. The following grouped data represent the rainfalls (in inches) from June to September in North-west India recorded for 83 years from 1875 to 1957 :

Rainfall	Frequency	Rainfall	Frequency
7-9	1	23-25	13
9-11	1	25-27	6
11-13	2	27-29	6
13-15	7	29-31	2
15-17	5	31-33	0
17-19	9	33-35	0
19-21	9	35-37	1
21-23	21	Total	83

The histogram of the data is shown below.

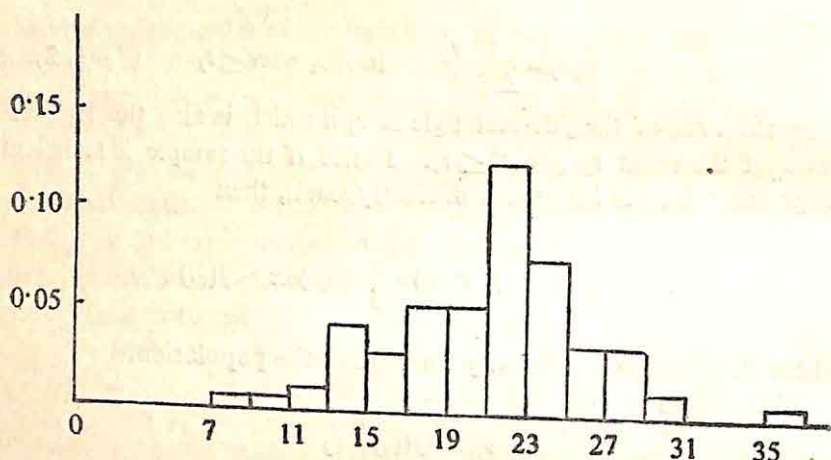


Fig. 24. Histogram for the Rainfall Data

12.4 SAMPLE CHARACTERISTICS

The characteristics such as the mean, variance, moments etc. of the empirical distribution of \hat{X} are called the *characteristics of the sample* or the *sample characteristics*. These are obviously functions of the sample values x_1, x_2, \dots, x_n and distinct from the characteristics of the population, i.e. of the random variable X . To differentiate these, we shall adopt the following useful system of notations. In general, a population characteristic will be denoted by a Greek letter, as already done in Ch. 7, and the corresponding sample characteristic by the corresponding Roman letter, except for some special cases, e.g. the population mean which is denoted by m . Accordingly, the principal sample characteristics are defined and denoted as follows.

sample mean = $E(\hat{X}) = \frac{1}{n} \sum x_i$, the arithmetic mean of the sample values which may be conveniently denoted by \bar{x} , i.e.

$$\bar{x} = \frac{1}{n} \sum x_i \quad (12.4.1)$$

The *sample variance* will be denoted by S^2 and defined by

$$S^2 = E\{(\hat{X} - \bar{x})^2\} = \frac{1}{n} \sum (x_i - \bar{x})^2 \quad (12.4.2)$$

The sample k th moment,

$$a_k = E(\dot{X}^k) = \frac{1}{n} \sum x_i^k \quad (12.4.3)$$

We have

$$a_0 = 1, a_1 = \bar{x}$$

The sample k th central moment,

$$m_k = E(\dot{X} - \bar{x})^k = \frac{1}{n} \sum (x_i - \bar{x})^k \quad (12.4.4)$$

So

$$m_0 = 1, m_1 = 0, m_2 = S^2$$

By (7.3.3) and (7.3.4) we have

$$m_k = \sum_{i=1}^k (-1)^i \binom{k}{i} a_{k-i} \bar{x}^i \quad (12.4.5)$$

or in details

$$\begin{aligned} m_2 &= S^2 = a_2 - \bar{x}^2 \\ m_3 &= a_3 - 3a_2\bar{x} + 2\bar{x}^3 \\ m_4 &= a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 3\bar{x}^4 \\ &\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \end{aligned} \quad (12.4.6)$$

According to (7.10.2) the *median* $z_{1/2}$ of the sample will be a point such that the number of sample values $< z_{1/2}$ is $\leq \frac{1}{2}n$, and the number of sample values $\leq z_{1/2}$ is $\geq \frac{1}{2}n$. Thus if n is odd, the median is always determinate. But if n is even, we have often a case in which every point of an interval between two sample values serves as a median; in such cases we follow our usual convention of taking the middle point of the interval as the proper sample median.

The *lower quartile* $z_{1/4}$ of the sample is a point such that the number of sample values $< z_{1/4}$ is $\leq \frac{1}{4}n$, and the number of sample values $\leq z_{1/4}$ is $\geq \frac{1}{4}n$. The *upper quartile* $z_{3/4}$ is similarly defined and the semi-interquartile range $= \frac{1}{2}(z_{3/4} - z_{1/4})$.

The *mode* of the sample is obviously the sample value having the maximum frequency, which will be denoted by \tilde{x} .

The mean, median and the mode are the important measures of location of the distribution of the sample. The important measures

of dispersion are the standard deviation and the semi-interquartile range ; another useful empirical measure of dispersion is the *range* of the sample which is defined to be the difference between the maximum and the minimum sample values.

$$\text{coefficient of skewness, } g_2 = \frac{m_3}{S^3} \quad (12.4.7)$$

$$\text{another measure of skewness} = \frac{\bar{x} - \tilde{x}}{S} \quad (12.4.8)$$

$$\text{coefficient of kurtosis, } b_2 = \frac{m_4}{S^4} \quad (12.4.9)$$

$$\text{coefficient of excess, } g_2 = \frac{m_4}{S^4} - 3 \quad (12.4.10)$$

12.5 COMPUTATION OF SAMPLE CHARACTERISTICS

Discrete population. When the population is discrete, we have have seen that the data is usually presented in a ξ - ν table, and we have

$$a_k = \frac{1}{n} \sum \nu_j \xi_j^k \quad (12.5.1)$$

The given two-column table is now extended by adding columns for $\nu \xi, \nu \xi^2, \dots, \nu \xi_j^k$, and the total in each column (except the first) is calculated. By (12.3.1) the total in the ν -column must be n , the size of the sample, and by dividing the other totals by n we obtain $a_1 = \bar{x}, a_2, \dots, a_k$. A final column representing $\nu(\xi+1)^k$ is often added for checking the computation by means of the following formula :

CHECK FORMULA

$$\sum \nu_j (\xi_j + 1)^k = \sum \nu_j \xi_j^k + \binom{k}{1} \sum \nu_j \xi_j^{k-1} + \binom{k}{2} \sum \nu_j \xi_j^{k-2} + \dots + \sum \nu_j \quad (12.5.2)$$

a_k 's being thus obtained, we can calculate the central moments m_k 's by (12.4.5) or (12.4.6). From these the important characteristics like the mean, variance, coefficients of skewness and excess etc. may be easily computed.

Linear transformation. We may sometimes considerably reduce our computational labour by making a linear transformation $\xi_j \rightarrow \xi_j'$ given by

$$\xi_j = a \xi_j' + b$$

which obviously amounts to the transformation $\overset{\circ}{X} = a\overset{\circ}{X}' + b$. If the characteristics of $\overset{\circ}{X}'$, which are first calculated, are marked with a prime, then it follows the theory of probability that

$$\bar{x} = a\bar{x}' + b, \quad S = |a|S', \quad m_k = a^k m_k' \quad (12.5.3)$$

and hence

$$g_1 = \frac{a}{|a|} g_1', \quad g_2 = g_2' \quad (12.5.4)$$

Example 1. Compute the mean, variance, mode, median, range and the semi-interquartile range of the sample given in Ex. 1 Sec. 12.3.

The computation is shown in the next page.

Set $\xi = \xi' + 7$.

ξ	v	ξ'	$v\xi'$	$v\xi'^2$	$v(\xi'+1)^2$
0	8	-7	- 56	392	288
1	59	-6	- 354	2124	1475
2	177	-5	- 885	4425	2832
3	311	-4	-1244	4976	2799
4	492	-3	-1476	4428	1968
5	528	-2	-1056	2112	528
6	601	-1	- 601	601	0
7	467	0	0	0	467
8	331	1	331	331	1324
9	220	2	440	880	1980
10	121	3	363	1089	1936
11	85	4	340	1360	2125
12	24	5	120	600	864
13	22	6	132	792	1078
14	6	7	42	294	384
15	3	8	24	192	243
Total	3455	—	-3880	24596	20291

The computation is first checked by formula (12.5.2).

$$\Sigma v\xi'^2 + 2\Sigma v\xi' + \Sigma v = 20291 = \Sigma v(\xi' + 1)^2$$

Dividing the totals in the 4th and 5th columns by $n = 3455$, we get

$$\bar{x} = -1.1230, a_2' = 7.1190$$

By (12.4.6)

$$S'^2 = 5.8579$$

The actual characteristics are now given by the transformation formulas (12.5.3).

$$\bar{x}' = 5.8770, S^2 = 5.8579$$

or retaining only upto 2 places of decimals

$$\bar{x} = 5.88, S^2 = 5.86$$

The mode, median, quartiles and range are obtain directly from the original table, i.e. the first two columns of the above table, the results being

$$\tilde{x} = 6, z_{1/2} = 6, z_{1/4} = 4, z_{3/4} = 7$$

so that the semi-interquartile range = 1.5. The maximum and minimum sample values are 15 and 0 respectively so that the range = 15.

Continuous population. For continuous populations, the data is usually grouped into classes, and as such the individual sample values x_1, x_2, \dots, x_n are not available for computation of the sample characteristics. We can, however, make approximate calculations by treating the class midpoints as ξ 's and the corresponding class frequencies as v 's of the discrete case. If the class intervals are sufficiently small, then this method will yield fairly good approximations to the actual characteristics. These approximations can also be sometimes slightly improved upon by what are known as Sheppard's corrections which we shall not, however, consider.

Example 2. Find the mean, standard deviation, coefficients of skewness and excess for the sample given in Ex. 2 Sec. 12.3.

$$\text{Set } \xi = 2\xi' + 22.$$

ξ	ν	ξ'	$\nu\xi'$	$\nu\xi'^2$	$\nu\xi'^3$	$\nu\xi'^4$	$\nu(\xi'+1)^4$
8	1	-7	-7	49	-343	2401	1296
10	1	-6	-6	36	-216	1296	625
12	2	-5	-10	50	-250	1250	512
14	7	-4	-28	112	-448	1792	567
16	5	-3	-15	45	-135	405	80
18	9	-2	-18	36	-72	144	9
20	9	-1	-9	9	-9	9	0
22	21	0	0	0	0	0	21
24	13	1	13	13	13	13	208
26	6	2	12	24	48	96	486
28	6	3	18	54	162	486	1536
30	2	4	8	32	128	512	1250
32	0	5	0	0	0	0	0
34	0	6	0	0	0	0	0
36	1	7	7	49	343	2401	4096
Total	83	—	-35	509	-779	10805	10686

$$\Sigma \nu \xi'^4 + 4 \Sigma \nu \xi'^3 + 6 \Sigma \nu \xi'^2 + 4 \Sigma \nu \xi' + \Sigma \nu = 10686 = \Sigma \nu (\xi' + 1)^4 \text{ (Checked)}$$

Now

$$\bar{x}' = -0.421687, \quad a_3' = 6.132530$$

$$a_5' = -9.385542, \quad a_4' = 130.180723$$

By (12.4.6)

$$S' = 2.440227, \quad m_3' = -1.777486, \quad m_4' = 120.797735$$

$$g_1' = -0.122325, \quad g_2' = 0.406729$$

Hence by (12.5.3)

$$\bar{x} = 21.156626, \quad S = 4.880454$$

$$g_1 = -0.122325, \quad g_2 = 0.406729$$

Finally, the results may be presented as

$$\bar{x} = 21.16, \quad S = 4.88, \quad g_1 = -0.12, \quad g_2 = 0.41$$

12.6 EXERCISES

1. The number of petals was counted for 22 flowers of a certain species with the following results :

4, 4, 7, 5, 4, 4, 4, 5, 6, 5, 6
9, 4, 4, 4, 4, 5, 6, 4, 5, 4, 4

Draw up a frequency table, and find the mean, median and mode of the sample.

2. The weekly wages in rupees of 25 workers of a factory were recorded to be

25.50, 21.00, 42.75, 31.75, 16.00
16.25, 20.25, 22.25, 24.75, 30.50
22.25, 20.25, 24.00, 36.75, 40.00
27.50, 23.00, 18.75, 20.25, 24.75
18.50, 33.00, 34.75, 20.25, 26.00

Arrange the data into suitable classes, and compute the mean and variance of the sample.

3. An experiment consists in throwing a die 5 times and noting the number of sixes. The experiment was repeated 200 times with the following results :

No. of sixes	0	1	2	3	4	5
Frequency	58	86	40	14	2	0

Find the sample mean and standard deviation.

4. The number of telephone calls received daily in a certain house in Calcutta was recorded for 92 days from 1st May to 31st July 1962, and the following data were obtained :

No. of calls	Frequency	No. of calls	Frequency
3	2	10	11
4	5	11	7
5	10	12	4
6	8	13	4
7	12	14	2
8	12		
9	15	Total	92

Compute the mean, variance, coefficients of skewness and excess, mode, median, semi-interquartile range and range of the sample.

5. The results of 150 determinations of the specific gravity of Ethyl Alcohol are classified as follows :

Specific gravity	Frequency	Specific gravity	Frequency
.765—.770	2	.795—.800	21
.770—.775	7	.800—.805	12
.775—.780	17	.805—.810	9
.780—.785	18	.810—.815	9
.785—.790	24	.815—.820	5
.790—.795	26	Total	150

Find the mean, standard deviation and the coefficients of skewness and excess of the sample.

SAMPLING DISTRIBUTIONS

13.1 SAMPLING DISTRIBUTIONS OF 'STATISTIC'S

Consider a sample x_1, x_2, \dots, x_n of size n drawn from the population of a given random variable X . We have remarked that the sample is a random sample in the sense that if we repeatedly draw samples of size n from the population of X under uniform conditions, the sets of sample values would fluctuate at random. This randomness of the sample may be mathematically described in the following way. The first sample value x_1 may be regarded as the observed value of a random variable X_1 , the second sample value x_2 that of another random variable X_2 etc., and finally the n th sample value the observed value of a random variable X_n . But all the sample values x_1, x_2, \dots, x_n are, in fact, observed values of the *parent* random variable X , and as such the random variables X_1, X_2, \dots, X_n must all have the same distribution, viz. that of X , i.e. of the population. Moreover, since the sample values are given by repetitions of the random experiment E under uniform conditions, it follows that the random variables X_1, X_2, \dots, X_n should be mutually independent. Note that these random variables are connected with the compound experiment of n independent repetitions of the given experiment E . To sum up, the sample values x_1, x_2, \dots, x_n are respectively regarded as observed values of the random variables X_1, X_2, \dots, X_n which are mutually independent, each having the distribution of the population. Or if we treat the sample values themselves as random variables, a random sample of size n from the population of X may be defined to be a set of n mutually independent and identically distributed random variables X_1, X_2, \dots, X_n , each having the distribution of X . This idea of random sample is of fundamental importance in mathematical statistics.

The n -dimensional random variable

$$\mathbf{x} = (X_1, X_2, \dots, X_n) \quad (13.1.1)$$

represents a random point in an n -dimensional space R^n , which will be called the *sample point* and R^n the *sample space*. A sample of size $n, (x_1, x_2, \dots, x_n)$ may then be regarded as an observed value of the

sample point x . The distribution function of the sample point, i.e. the joint distribution function of X_1, X_2, \dots, X_n is given by

$$F(x_1, x_2, \dots, x_n) = F(x_1) F(x_2) \dots F(x_n) \quad (13.1.2)$$

where $F(x)$ denotes the distribution function of the population.

Any function of the sample values x_1, x_2, \dots, x_n is, in general, called a *statistic*; for example, the sample mean \bar{x} , the sample variance S^2 etc. are all statistics. Let $a = a(x_1, x_2, \dots, x_n)$ denote any statistic which may be regarded as an observed value of the corresponding random variable $A = a(X_1, X_2, \dots, X_n)$. Now given the distribution function of the population $F(x)$, the joint distribution of X_1, X_2, \dots, X_n is obtained from (13.1.2) which, we know, can uniquely determine the distribution of the function $A = a(X_1, X_2, \dots, X_n)$ of the random variables X_1, X_2, \dots, X_n . This probability distribution of the random variable A will be called the *sampling distribution* of the statistic $a = a(x_1, x_2, \dots, x_n)$.

Notations. For avoiding the boredom of frequent restatements, we shall stick to the following permanent system of notations as far as practicable.

(i) x_1, x_2, \dots, x_n will denote a sample of size n from the population of X , which will be treated as observed values of the random variables X_1, X_2, \dots, X_n respectively. The running real variables corresponding to X_1, X_2, \dots, X_n , for writing the distribution function etc., will be denoted, in keeping with our usual practice in the theory of probability, again by x_1, x_2, \dots, x_n as we have already done in (13.1.2). We hope that this will not give rise to any confusion; it will always be clear from the context if x_1, x_2, \dots, x_n denote the sample values or the real variables.

(ii) A statistic will usually (with some exceptions) be denoted by a small Roman letter, as we have proposed for the case of sample characteristics; the corresponding random variable will then be denoted by the corresponding bold capital latter, e.g. if $a = a(x_1, x_2, \dots, x_n)$ is any statistic, the random variable corresponding to it will be $A = a(X_1, X_2, \dots, X_n)$.

In the above notations, the random variables corresponding to the sample characteristics are given by

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum X_i & S^2 &= \frac{1}{n} \sum (X_i - \bar{X})^2 \\ A_k &= \frac{1}{n} \sum X_i^k, & M_k &= \frac{1}{n} \sum (X_i - \bar{X})^k \\ G_1 &= \frac{M_3}{S^3}, & G_2 &= \frac{M_4}{S^4} - 3\end{aligned} \quad (13.1.3)$$

etc. Speaking about the sample characteristics, we may also conveniently say—sampling distributions of the mean, variance etc. to mean the probability distributions of \bar{X} , S^2 etc. respectively.

When a sample of size n is drawn from the population of X , we get an observed value of the n -dimensional random variable x , the sample point, so that we can also conceive of the population of the sample point if we fix our basic random experiment as drawing a sample of size n from the population of X , which, in other words, is the compound experiment of n independent trials of the given experiment E . If a sequence of m samples of fixed size n are drawn under uniform conditions, we shall get a sample of size m from the population of x . Likewise, a sample of size n from the population of X yields an observed value of any statistic $A = a(X_1, X_2, \dots, X_n)$, and consequently an infinite sequence of independent samples (i.e. samples drawn under uniform conditions) will give rise to the population of the statistic A , from which a sample of any finite size may be drawn. We note that any particular observed value $a = a(x_1, x_2, \dots, x_n)$ of A can be regarded as a sample of unit size from the population of A .

13.2. ESTIMATES—CONSISTENT AND UNBIASED

Let a be an unknown *characteristic* or *parameter* of the population. (It may sometimes so happen that the population distribution function $F(x)$ has a known mathematical form which contains a number of unknown parameters, e.g. the population is known to be normal (m, σ) where m, σ are unknown parameters; a may also denote any such parameter.) In order to determine an experimental values of a on the basis of a sample, perhaps a natural suggestion is to find a statistic $a = a(x_1, x_2, \dots, x_n)$ whose computed value is approximately equal to a .

This is, however, a very unprecise statement and means little mathematically, for we know that the computed value of a statistic fluctuates at random from sample to sample, and if the computed value of a is close to a for one sample, it may deviate considerably from a for another sample. Thus it would be meaningless to infer anything from a particular observed value of the statistic. A proper judgment can, however, be obtained from the sampling distribution of the statistic. If the probability mass in the sampling distribution of the statistic a is concentrated near the point a , then we may say that it is highly probable that an observed value of a will lie in a given small neighbourhood of a . In this sense the statistic a will be called an *estimate* of the population characteristic or parameter a . A measure of *goodness* or *precision* of the estimate is naturally given by a measure of concentration or inversely by a measure of dispersion of the sampling distribution of the statistic a about the point a , and a useful measure of this dispersion is furnished by $E\{(A - a)^2\}$ or its positive square root. The above definition of an estimate is very general and somewhat loose, and accordingly, we may have different estimates of the same population characteristic or parameter with varying degrees of precision. An estimate a_1 of a will be said to be better than another estimate a_2 if the sampling distribution of a_1 is more concentrated about a than that of a_2 , or if $E\{(A_1 - a)^2\} < E\{(A_2 - a)^2\}$.

There are several desirable types of estimates, of which we shall consider only two, viz. consistent and unbiased estimates.

A statistic $a = a(x_1, x_2, \dots, x_n)$ is said to be a *consistent estimate* of a population characteristic or parameter a if

$$A \xrightarrow[\text{in } p]{\quad} a \quad \text{as } n \rightarrow \infty \quad (13.2.1)$$

i.e. the precision of the estimate increases with the size of the sample, and hence such an estimate is expected to give very accurate results for large samples. The property of consistency is obviously a desirable property for good estimates.

A statistic $a = a(x_1, x_2, \dots, x_n)$ will be called an *unbiased estimate* of a population characteristic or parameter a if

$$E(A) = a \quad (13.2.2)$$

In case an estimate a is such that $E(A) \neq a$, then it is said to be *biased*. The quantity $E(A) - a$ is called the *bias* of the estimate a ; the estimate is said to be *positively* or *negatively biased* according as the bias is positive or negative. In view of the importance of the mean as a parameter of location in a probability distribution, the property of unbiasedness immediately recommends itself for good estimates. For an unbiased estimate a , a convenient inverse measure of precision is provided by $\sigma(A)$.

Remark. The concepts of consistency and unbiasedness are independent, i.e. one does not imply the other. A consistent estimate may be biased and an unbiased estimate inconsistent. The bias of a consistent estimate, however, decreases with increasing size of the sample. An estimate which is only unbiased is not necessarily good, for it refers nothing to the precision of the estimate. Estimates which are both consistent and unbiased can certainly be regarded as very good estimates.

In the previous chapter, we showed that the empirical distribution of the sample is a statistical image of the population distribution, and we must have been wondering what would be the connections between the sample characteristics and those of the population. These can now be precisely stated as the following theorems.

Theorem I. a_k is a consistent and unbiased estimate of a_k , provided the latter exists.

Proof. By (13.1.3) $A_k = \frac{1}{n} \sum X_i^k$ and for all i

$$E(X_i^k) = E(X^k) = a_k$$

Since X_1, X_2, \dots, X_n are mutually independent random variables all having the population distribution, $X_1^k, X_2^k, \dots, X_n^k$ are also mutually independent having a common distribution with mean a_k , and hence if a_k exists, it follows from the Law of Large Numbers for equal components that

$$A_k \xrightarrow[\text{in } p]{} a_k \quad \text{as } n \rightarrow \infty$$

i.e. a_k is a consistent estimate of a_k .

Moreover

$$E(A_k) = \frac{1}{n} \sum E(X_i^k) = \alpha_k$$

which shows that the estimate is also unbiased.

Theorem II. If μ_k exists, m_k is a consistent estimate of μ_k .

Proof. From (12.4.5)

$$M_k = \sum_{i=0}^k (-1)^i \binom{k}{i} A_{k-i} \bar{X}^i$$

If μ_k exists, all lower order moments also exist, and hence by Theorem I

$$M_k \xrightarrow{\text{in } p} \sum_{i=0}^k (-1)^i \binom{k}{i} \alpha_{k-i} m^i = \mu_k \quad \text{as } n \rightarrow \infty$$

This proves the theorem. We remark that m_k 's are not always unbiased estimates of μ_k 's as we shall presently see.

COROLLARY. g_1 and g_2 are consistent estimates of γ_1 and γ_2 respectively.

Thus, speaking broadly, for large samples the characteristics of the sample give approximate values of the corresponding characteristics of the population.

We know that the sampling distributions of the characteristics are uniquely determined by the distribution function of the population $F(x)$. But the actual determinations of the distribution functions of the sampling distributions in terms of $F(x)$ often present great mathematical difficulties, and general analytical formulas for these are mostly unknown. In the following section, we shall study some properties of the sampling distributions, and work out exact results for a particularly simple and important population, viz. the normal population in the next.

13.3 IMPORTANT SAMPLING DISTRIBUTIONS

Sample mean

$$\bar{X} = \frac{1}{n} \sum X_i$$

Since X_1, X_2, \dots, X_n are mutually independent each having the distribution of the population, we have by (8.5.6) and (8.5.8)

$$E(\bar{X}) = m, \quad \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} \quad (13.3.1)$$

assuming that the population mean and standard deviation exist. By Theorem I the sample mean \bar{x} is a consistent and unbiased estimate of the population mean m , an inverse measure of precision of the estimate being given by $\sigma(\bar{X}) = \sigma/\sqrt{n}$ which decreases as n increases.

Also it follows from the Central Limit Theorem for equal components that if σ exists, \bar{X} is asymptotically normal $(m, \sigma/\sqrt{n})$.

Sample variance, Unbiased estimate of the population variance

$$S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

We may write

$$S^2 = \frac{1}{n} \sum (X_i - m)^2 - (\bar{X} - m)^2$$

So

$$\begin{aligned} E(S)^2 &= \frac{1}{n} \sum E\{(X_i - m)^2\} - E\{(\bar{X} - m)^2\} \\ &= \sigma^2 - \frac{\sigma^2}{n} \end{aligned} \quad [\text{by (13.3.1)}]$$

or

$$E(S^2) = \frac{n-1}{n} \sigma^2 \quad (13.3.2)$$

We know from Theorem II that S^2 is a consistent estimate of σ^2 ; but (13.3.2) shows that the estimate is not unbiased. Since $E(S^2) < \sigma^2$, the bias in this case is negative. Now if we set

$$s^2 = \frac{n}{n-1} S^2 \quad (13.3.3)$$

$$s^2 \xrightarrow{\text{in } p} \sigma^2 \text{ as } n \rightarrow \infty, \quad E(s^2) = \sigma^2$$

i.e. s^2 is a consistent as well as an unbiased estimate of σ^2 ; s^2 will be referred to as the *unbiased estimate of the population variance*.

If σ is known, we can calculate the value of σ/\sqrt{n} which gives an inverse measure of precision of the estimate \bar{x} of the population mean m . But if σ is unknown, as is usually the case, this cannot be done. In that case an approximate value of the same may be obtained by replacing σ by a good estimate say, S or s , i.e. S/\sqrt{n} or s/\sqrt{n} may be taken to be an approximate inverse measure of precision of the estimate \bar{x} of m .

13.4 NORMAL POPULATION

It has been found from experience that a strikingly large variety of populations met in practice have normal or approximately normal distributions, e.g. populations of heights, weights etc. of racially homogeneous people, temperatures, rainfalls etc. for a season, experimentally measured values of a physical quantity, marks obtained in an examination and so on. This may seem somewhat strange at first sight but can be largely accounted for by the unique position of the normal distribution offered by the Central Limit Theorem. In many cases (a significant example of which will be found in the theory of errors) the random variable in question may be conceived as the sum of a large number of independent random variables arising out of a large number of random causes, and hence is approximately normally distributed by virtue of the Central Limit Theorem. This is why the normal distribution assumes great importance in statistics. Luckily for us the calculations with the normal distribution are also comparatively easy, and we shall be able to find the exact forms of the sample distributions of the mean, variance etc. in the case of the normal population.

Consider a normal (m, σ) population. Then X_1, X_2, \dots, X_n is a set of n mutually independent variates, each normal (m, σ) , and the following theorems hold.

Theorem I. The sample mean \bar{X} is normal $(m, \sigma/\sqrt{n})$.

Proof. Observing that \bar{X} is a linear combination of X_1, X_2, \dots, X_n , the theorem follows as a particular case of the reproductive property of the normal distribution (cf. Sec. 8.8).

COROLLARY. The statistic $U = \frac{\sqrt{n}(\bar{X} - m)}{\sigma}$ is normal $(0, 1)$.

Theorem II. The statistic $\chi^2 = \frac{nS^2}{\sigma^2}$ has a χ^2 -distribution with $\nu = n - 1$ degrees of freedom, and the sample mean \bar{X} and sample variance S^2 are independent variates.

Proof. This theorem is a simple consequence of Theorem III Sec. 9.1. We have

$$S^2 = \frac{1}{n} \sum (X_i - m)^2 - (\bar{X} - m)^2$$

or

$$\chi^2 = \frac{nS^2}{\sigma^2} = \sum \left(\frac{X_i - m}{\sigma} \right)^2 - \left\{ \frac{\sqrt{n}(\bar{X} - m)}{\sigma} \right\}^2$$

Now $\frac{X_1 - m}{\sigma}, \frac{X_2 - m}{\sigma}, \dots, \frac{X_n - m}{\sigma}$ are n mutually independent standard normal variates, and

$$\frac{\sqrt{n}(\bar{X} - m)}{\sigma} = \frac{1}{\sqrt{n}} \left(\frac{X_1 - m}{\sigma} \right) + \frac{1}{\sqrt{n}} \left(\frac{X_2 - m}{\sigma} \right) + \dots + \frac{1}{\sqrt{n}} \left(\frac{X_n - m}{\sigma} \right)$$

is a linear combination of them such that

$$\left(\frac{1}{\sqrt{n}} \right)^2 + \left(\frac{1}{\sqrt{n}} \right)^2 + \dots + \left(\frac{1}{\sqrt{n}} \right)^2 = 1$$

Hence, by Theorem III Sec. 9.1, χ^2 is χ^2 -distributed with $\nu = n - 1$ degrees of freedom and χ^2 is independent of $U = \sqrt{n}(\bar{X} - m)/\sigma$ so that \bar{X} and S^2 are independent.

Distributions of S^2 and s^2 . With the help of the above theorem, we may now easily determine the density functions for S^2 and s^2 . The probability differential for χ^2

$$dF = \frac{e^{-\frac{1}{2}\chi^2} \left(\frac{1}{2}\chi^2 \right)^{\nu/2-1}}{2\Gamma\left(\frac{1}{2}\nu\right)} d\chi^2 \quad [\nu = n - 1]$$

$$= \frac{1}{\Gamma\left(\frac{1}{2}(n-1)\right)} \left(\frac{n}{2\sigma^2} \right)^{(n-1)/2} (S^2)^{(n-3)/2} e^{-nS^2/2\sigma^2} dS^2$$

So

$$f_{S^2}(S^2) = \frac{1}{\Gamma\left(\frac{1}{2}(n-1)\right)} \left(\frac{n}{2\sigma^2} \right)^{(n-1)/2} (S^2)^{(n-3)/2} e^{-nS^2/2\sigma^2}$$

$$(0 < S^2 < \infty) \quad (13.4.1)$$

Similarly, noting that $\chi^2 = \nu s^2 / \sigma^2$, we have

$$f_{s^2}(s^2) = \frac{1}{\Gamma(\frac{1}{2}\nu)} \left(\frac{\nu}{2\sigma^2} \right)^{\nu/2} (s^2)^{\nu/2-1} e^{-\nu s^2/2\sigma^2} \quad (0 < s^2 < \infty) \quad (13.4.2)$$

Theorem III. The statistic $t = \frac{\sqrt{n}(\bar{X} - m)}{s}$, known as *Student's ratio*, is t -distributed with $\nu = n - 1$ degrees of freedom.

Proof. Since $\chi^2 = \nu s^2 / \sigma^2$, we may write $t = \sqrt{\nu} U / \sqrt{\chi^2}$ where U is normal $(0, 1)$, χ^2 is χ^2 -distributed with ν degrees of freedom, and U and χ^2 are independent. Hence the above theorem follows from Theorem I Sec. 9.2.

Remark. The statistics U , χ^2 and t introduced above will be useful in the theories of estimation and testing of hypotheses for the normal population. We note that the distribution of each of these statistics is independent of the population parameters m, σ ; but U depends on both m and σ , χ^2 only on σ and t only on m .

13.5 EXERCISES

1. Show that the coefficients of skewness and excess of the sampling distribution of the mean are respectively γ_1 / \sqrt{n} and γ_2 / n , n being the size of the sample and γ_1 and γ_2 the corresponding coefficients of the population.

2. Show that $\sigma^2(s^2)$ which serves as an inverse measure of precision of s^2 , the unbiased estimate of the population variance σ^2 , is given by

$$\sigma^2(s^2) = \frac{1}{n} \left\{ \mu_4 - \frac{n-3}{n-1} \sigma^4 \right\}$$

3. Prove the formulas

$$E(M_3) = \frac{(n-1)(n-2)}{n^2} \mu_3$$

$$E(M_4) = \frac{(n-1)(n^2-3n+3)}{n^3} \mu_4 + \frac{3(n-1)(2n-3)}{n^3} \sigma^4$$

Hence deduce that

$$\frac{n^2 m_3}{(n-1)(n-2)} \text{ and } \frac{n^2}{(n-1)(n-2)(n-3)} [(n+1)m_4 - 3(n-1)S^4]$$

are consistent and unbiased estimates of the cumulants κ_3 and κ_4 respectively. Also show that the corrected estimates of γ_1 and γ_2 for small samples are respectively

$$\frac{\sqrt{n(n-1)}}{n-2} g_1 \text{ and } \frac{n-1}{(n-2)(n-3)} [(n+1)g_2 + 6]$$

4. Show that A_k is asymptotically normal $(a_k, \sqrt{(a_{2k} - a_k^2)/n})$ if a_{2k} exists.
5. Using the statistic χ^2 , prove that for a normal (m, σ) population

$$\sigma(s^2) = \sqrt{2/(n-1)} \sigma^2$$

Verify the result from the general formula in Ex. 2. Hence show that an estimate of $\sigma(s^2)$ is $\sqrt{2/(n-1)} s^2$.

6. Find the sampling distribution of the mean for the (a) binomial, (b) Poisson and (c) gamma populations.

7. Show that the sample mean \bar{X} and the sample variance S^2 are uncorrelated if $\mu_3 = 0$.

ESTIMATION OF PARAMETERS

14.1 METHOD OF MAXIMUM LIKELIHOOD

Let us suppose that the distribution function of the population, $F(x)$ has a known functional form but contains a number of unknown parameters $\theta_1, \theta_2, \dots, \theta_k$, and our problem is to find estimates of these parameters on the basis of a sample: x_1, x_2, \dots, x_n drawn from the population. There are several methods by which such estimation can be done, of which the most important is the method of *maximum likelihood*. The importance of this method lies in the fact that in most cases it yields very good estimates. These estimates are found to be good by many yardsticks, but we do not here propose to enter into the mathematical discussions of the same.

In this method our first task is to define what is called the *likelihood function* of the sample. This is done separately for discrete and continuous populations as follows.

DISCRETE CASE. Let X denote the parent random variable, and, for convenience, let us write

$$P(X=x_i) = f_{x_i}(\theta_1, \theta_2, \dots, \theta_k) \quad (14.1.1)$$

The event that the particular sample x_1, x_2, \dots, x_n has been drawn is $(X_1=x_1, X_2=x_2, \dots, X_n=x_n)$, and the probability of this event, which is clearly a function of the sample values x_1, x_2, \dots, x_n and the parameters $\theta_1, \theta_2, \dots, \theta_k$, is defined to be the likelihood function of the sample to be denoted by $L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k)$, i.e.

$$L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = P(X_1=x_1, X_2=x_2, \dots, X_n=x_n)$$

Now since X_1, X_2, \dots, X_n are mutually independent each having the distribution of X , we have

$$L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = f_{x_1}(\theta_1, \theta_2, \dots, \theta_k) f_{x_2}(\theta_1, \theta_2, \dots, \theta_k) \dots f_{x_n}(\theta_1, \theta_2, \dots, \theta_k) \quad (14.1.2)$$

CONTINUOUS CASE. In this case the event of drawing the particular sample may be represented by $(x_1 < X_1 \leq x_1 + dx_1, x_2 < X_2 \leq$

$x_2 + dx_2, \dots, x_n < X_n \leq x_n + dx_n$, the probability of which is obviously the probability differential of the sample point x , $f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$ where $f(x_1, x_2, \dots, x_n)$ is the density function of x . The density function of x will be, in the continuous case, defined to be the likelihood function, i.e.

$$L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = f(x_1, x_2, \dots, x_n)$$

Hence if $f(x; \theta_1, \theta_2, \dots, \theta_k)$ denotes the density function of the population, then

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) \\ = f(x_1; \theta_1, \theta_2, \dots, \theta_k) f(x_2; \theta_1, \theta_2, \dots, \theta_k) \dots f(x_n; \theta_1, \theta_2, \dots, \theta_k) \end{aligned} \quad (14.1.3)$$

When the sample values are regarded as *fixed*, the likelihood function L becomes a function of the parameters $\theta_1, \theta_2, \dots, \theta_k$ only, and the method of maximum likelihood consists in finding those values of the parameters as functions of x_1, x_2, \dots, x_n which would maximise the likelihood function. Thus if the function L has a unique maximum for

$$\theta_1 = \hat{\theta}_1(x_1, x_2, \dots, x_n), \theta_2 = \hat{\theta}_2(x_1, x_2, \dots, x_n), \dots, \theta_k = \hat{\theta}_k(x_1, x_2, \dots, x_n)$$

then the statistics $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ are called the *maximum likelihood estimates* of $\theta_1, \theta_2, \dots, \theta_k$ respectively.

Since $L > 0$, maximising L amounts to maximising $\log L$, the equations for which are

$$\frac{\partial \log L}{\partial \theta_1} = 0, \frac{\partial \log L}{\partial \theta_2} = 0, \dots, \frac{\partial \log L}{\partial \theta_k} = 0 \quad (14.1.4)$$

These are called the *likelihood equations*, by solving which we can find the maximum likelihood estimates of $\theta_1, \theta_2, \dots, \theta_k$, provided they exist.

14.2 APPLICATIONS TO DIFFERENT POPULATIONS

1. Binomial (N, p) population. Of the two parameters N is usually known, and our problem remains to estimate the parameter p . Here

$$f_{x_i}(p) = \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$$

By (14.1.2)

$$L = \binom{N}{x_1} \binom{N}{x_2} \cdots \binom{N}{x_n} p^{x_1+x_2+\cdots+x_n} (1-p)^{nN-(x_1+x_2+\cdots+x_n)}$$

So

$$\log L = (x_1 + x_2 + \cdots + x_n) \log p + \{nN - (x_1 + x_2 + \cdots + x_n)\} \log(1-p) + \text{terms independent of } p$$

The likelihood equation is $\frac{\partial \log L}{\partial p} = 0$ which gives

$$\frac{x_1 + x_2 + \cdots + x_n}{p} = \frac{nN - (x_1 + x_2 + \cdots + x_n)}{1-p} = nN$$

or

$$p = (x_1 + x_2 + \cdots + x_n)/nN = \bar{x}/N$$

i.e.

$$\hat{p} = \bar{x}/N$$

We know that the sample mean \bar{x} is a consistent and unbiased estimate of the population mean which, in this case, is Np , and hence it follows that \hat{p} is a consistent and unbiased estimate of p .

If, instead of a sample of size n , we consider a single observed value v of the parent variable X , which may be regarded as a sample of size 1, then we have

$$\hat{p} = v/N$$

The interpretation of the above result in terms of a Bernoullian sequence of trials appears very plausible. We know that the number of successes in a Bernoullian sequence of N trials with probability of success p is binomial (N, p) , and the above result states that the maximum likelihood estimate of the probability of success is equal to the observed value of the frequency ratio of the same.

Remark. We have $E(X/N) = p$, and Bernoulli's theorem states that $X/N \xrightarrow{\text{in } p} p$ as $N \rightarrow \infty$. These show that, for large N , $\hat{p} = v/N$ is a good estimate of the parameter p .

2. Poisson- μ population

$$f_{x_i}(\mu) = e^{-\mu} \frac{\mu^{x_i}}{x_i!}$$

So

$$L = e^{-n\mu} \frac{\mu^{n\bar{x}}}{x_1! x_2! \dots x_n!}$$

or

$$\log L = -n\mu + n\bar{x} \log \mu + \text{terms independent of } \mu$$

The equation $\frac{\partial \log L}{\partial \mu} = 0$ gives $\mu = \bar{x}$ or $\hat{\mu} = \bar{x}$.

Since μ is the mean of the population, its maximum likelihood estimate $\hat{\mu}$ is both consistent and unbiased.

3. Normal (m, σ) population

$$f(x; m, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-m)^2/2\sigma^2}$$

By (14.1.3)

$$L = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m)^2}$$

So

$$\log L = -\frac{1}{2}n \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum (x_i - m)^2$$

The likelihood equations are

$$\frac{\partial \log L}{\partial m} = 0, \quad \frac{\partial \log L}{\partial \sigma} = 0$$

The first equation gives $\sum (x_i - m) = 0$ or $m = \bar{x}$, and the second

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (x_i - m)^2 = 0$$

or

$$\sigma^2 = \frac{1}{n} \sum (x_i - m)^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = S^2$$

Hence $\hat{m} = \bar{x}$, sample mean, and $\hat{\sigma}^2 = S^2$, the sample variance or $\hat{\sigma} = S$. We know that the estimate \hat{m} is consistent and unbiased, whereas the estimate $\hat{\sigma}^2$ is consistent but biased.

Remarks

1. For a given set of parameters, we can construct different likelihood functions which may produce different estimates of the same parameters. Let us take the following example.

For estimating the parameter σ of the normal population, we may, instead of the parent population, consider the population of the statistic s^2 whose density function is given by (13.4.2). From the given sample x_1, x_2, \dots, x_n the value of s^2 is calculated which forms a sample of size 1 from the population of s^2 , the likelihood function for which is given by

$$L(s^2; \sigma) = f_{s^2}(s^2; \sigma) = \frac{1}{\Gamma(\frac{1}{2}v)} \left(\frac{v}{2\sigma^2} \right)^{v/2} (s^2)^{v/2-1} e^{-vs^2/2\sigma^2}$$

So

$$\log L = -v \log \sigma - \frac{vs^2}{2\sigma^2} + \text{terms independent of } \sigma$$

The likelihood equation $\frac{\partial \log L}{\partial \sigma} = 0$ gives $\hat{\sigma}^2 = s^2$ which is a consistent and unbiased estimate of σ^2 .

2. In all the above examples we, may easily verify that the estimates obtained actually correspond to a unique maximum of the likelihood function.

14.3 INTERVAL ESTIMATION

We have so far estimated a population parameter by means of a single statistic. Such estimation by a single statistic is called *point estimation*, and a point estimate when computed from an observed sample is supposed to give a value somewhat close to the true value of the estimated parameter. This, however, provides very insufficient information about the true value of the parameter unless some measure of goodness or precision of the estimate is also given. We, of course, know that a useful inverse measure of precision or a measure of *uncertainty* of an estimate a of a parameter α is given by $\sqrt{E\{(A - \alpha)^2\}}$ or the like. (In case the latter contains unknown population parameters, we may calculate an approximate value of the same by replacing the parameters by their estimates from the sample). Thus given such a measure of uncertainty, we can indeed

compare the goodness of different estimates of the same parameter, but it still remains largely unknown how to make exact use of this measure as an *error* or *correction* of the estimate. It was found that this problem can only be satisfactorily tackled by the method of *interval estimation* which makes use of a pair of statistics forming an interval. The idea of interval estimation may be precisely stated as follows.

Let α be a population parameter and ε ($0 < \varepsilon < 1$) a given number. If there exist two statistics

$$a = a(x_1, x_2, \dots, x_n) \quad \text{and} \quad b = b(x_1, x_2, \dots, x_n)$$

such that

$$P(A < \alpha < B) = 1 - \varepsilon \quad (14.3.1)$$

where $A = a(X_1, X_2, \dots, X_n)$ and $B = b(X_1, X_2, \dots, X_n)$ are the random variables corresponding to the statistics a and b respectively, then the interval (a, b) is called an *interval estimate* or a *confidence interval* for the parameter α with confidence coefficient $1 - \varepsilon$; the statistics a, b are respectively called the *lower* and *upper confidence limits* for α .

The probability statement of (14.3.1) is somewhat queer, a like of which did not appear anywhere in the theory of probability, but nevertheless has a definite meaning. We note that here A and B are random variables but α is a fixed constant, so that (14.3.1) states that the probability that the random interval (A, B) covers the point α is $1 - \varepsilon$. A practical interpretation of this will be that if a long sequence of random samples are drawn under uniform conditions and the statistics a, b computed each time, then the ratio of the number of times the interval (a, b) includes the true parameter value α to the total number of samples drawn is approximately equal to $1 - \varepsilon$. The number ε is usually chosen to be small, say, .05 or .01 or .001 etc., i.e. the confidence coefficient is .95 or .99 or .999 etc., the corresponding confidence interval being then called a 95% or 99% or 99.9% etc. confidence interval. For a 95% confidence interval, we may roughly say that in repeatedly asserting that the parameter lies in the confidence interval for a large number of samples, we are liable to a risk of error only in 5% of the cases.

It is also possible to find many confidence intervals for a parameter corresponding to a given sample and a given confidence coefficient. To compare the relative goodness of these, we can use the length of the interval, $b - a$ as an *inverse measure of precision* of the interval estimate; of two confidence intervals, the one having the smaller length is obviously preferable.

For a given sample the length of a confidence interval depends on ε ; the dependence, in general, is such that as ε decreases, i.e. the confidence coefficient increases, the length of the interval also increases making the estimate worse and worse. Thus in order to have a very accurate estimate, we must agree to low value of the confidence coefficient, which is, however, not very useful from the practical point of view. Hence for practical problems we must strike an optimum between the level of confidence and the precision of the interval estimate. As remarked earlier 95%, 99% etc. confidence intervals are frequently used in practice.

14.4 METHOD FOR FINDING CONFIDENCE INTERVALS

We shall here outline a method for obtaining confidence intervals which, although not perfectly general, is quite useful in many important cases. For convenience, only continuous populations will be considered in the sequel.

Let $\theta_1, \theta_2, \dots, \theta_k$ be the unknown population parameters, of which we want to estimate, say, θ_1 .

1. Choose, if possible, a statistic

$$z = z(x_1, x_2, \dots, x_n; \theta_1) \quad (14.4.1)$$

whose sampling distribution is independent of *all* the parameters and which itself depends on θ_1 but independent of $\theta_2, \theta_3, \dots, \theta_k$; these unwanted parameters $\theta_2, \theta_3, \dots, \theta_k$ are often called *nuisance parameters*.

2. Now choose two numbers $\alpha_\varepsilon, \beta_\varepsilon (> \alpha_\varepsilon)$ depending on ε such that

$$\int_{\alpha_\varepsilon}^{\beta_\varepsilon} f_z(z) dz = 1 - \varepsilon \quad (14.4.2)$$

where $f_z(z)$ is the density function of Z , which is independent of all unknown parameters. This can usually be done in infinitely many ways which would lead to infinitely many confidence intervals.

3. Eq. (14.4.2) states that

$$P(\alpha_\epsilon < Z < \beta_\epsilon) = 1 - \epsilon \quad (14.4.3)$$

Now if the statistic z is such a function of θ_1 that the inequalities $\alpha_\epsilon < Z < \beta_\epsilon$ can be inverted to the form $A < \theta_1 < B$ where A and B are random variables corresponding to the statistics a and b respectively which depend on ϵ , then (14.4.3) can be re-written as

$$P(A < \theta_1 < B) = 1 - \epsilon$$

This shows that (a, b) is a desired confidence interval for θ_1 having confidence coefficient $1 - \epsilon$.

Remark. It is often difficult to find a suitable statistic z as described above, and this curtails the generality of the method to a great extent.

14.5 APPLICATIONS TO NORMAL (m, σ) POPULATION

Confidence interval for m

Case I. σ known. We can here conveniently choose the statistic

$$u = \frac{\sqrt{n}(\bar{x} - m)}{\sigma}$$

whose sampling distribution is normal $(0, 1)$ and which depends on m , the parameter to be estimated.

Take two points $\pm u_\epsilon$ symmetrically about the origin such that

$$P(-u_\epsilon < U < u_\epsilon) = 1 - \epsilon$$

or

$$P\left(-u_\epsilon < \frac{\sqrt{n}(\bar{X} - m)}{\sigma} < u_\epsilon\right) = 1 - \epsilon$$

which can be re-written as

$$P\left(\bar{X} - \frac{\sigma u_\epsilon}{\sqrt{n}} < m < \bar{X} + \frac{\sigma u_\epsilon}{\sqrt{n}}\right) = 1 - \epsilon$$

Hence a confidence interval for m having confidence coefficient $1 - \epsilon$ is

$$\left(\bar{x} - \frac{\sigma u_\epsilon}{\sqrt{n}}, \bar{x} + \frac{\sigma u_\epsilon}{\sqrt{n}} \right) \quad (14.5.1)$$

where u_ϵ is given by

$$P(-u_\epsilon < U < u_\epsilon) = 1 - \epsilon$$

or, from the symmetry of the standard normal distribution, by

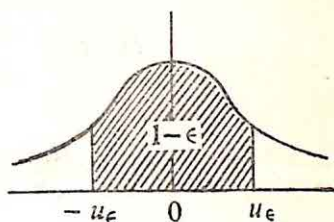


Fig. 25

$$P(U > u_\epsilon) = \frac{1}{2}\epsilon \quad (14.5.2)$$

Remark. Instead of choosing the symmetrical points $\pm u_\epsilon$, we may also choose any two points unsymmetrically and obtain the corresponding confidence interval by an exactly similar method. But it can be proved that among all these the symmetrical points lead to the shortest and hence the best interval.

Example 1. The mean of a sample of size 50 from a normal population is found to be 15.68. If it is known that the standard deviation of the population is 3.27, find 95% and 99% confidence intervals for the population mean.

For $\epsilon = .05$, u_ϵ is given by

$$P(U > u_\epsilon) = .025$$

From Table I at the end of the book, $u_\epsilon = 1.960$.

Now $n = 50$, $\bar{x} = 15.68$, $\sigma = 3.27$, so that the computed value of the confidence interval (14.5.1) becomes (14.77, 16.59). This is a 95% interval for the population mean.

Similarly, a 99% confidence interval is (14.49, 16.87).

The length of the 95% interval is 1.82, whereas that of the 99% interval is 2.38.

Case II. σ unknown. Here σ is a nuisance parameter, and the statistic u is no longer applicable as it contains σ . In this case our choice falls on Student's ratio

$$t = \frac{\sqrt{n}(\bar{x} - m)}{s}$$

the sampling distribution of which, we know, is t -distributed with $\nu = n - 1$ degrees of freedom.

Determine two numbers $\pm t_\epsilon^m$ by

$$P(-t_\epsilon < t < t_\epsilon) = 1 - \epsilon$$

or

$$P(-t_\epsilon < \frac{\sqrt{n}(\bar{X} - m)}{S} < t_\epsilon) = 1 - \epsilon$$

or

$$P\left(\bar{X} - \frac{St_\epsilon}{\sqrt{n}} < m < \bar{X} + \frac{St_\epsilon}{\sqrt{n}}\right) = 1 - \epsilon$$

which shows that

$$\left(\bar{x} - \frac{st_\epsilon}{\sqrt{n}}, \bar{x} + \frac{st_\epsilon}{\sqrt{n}}\right) \quad (14.5.3)$$

is a required confidence interval for m .

Here t_ϵ is given by

$$P(-t_\epsilon < t < t_\epsilon) = 1 - \epsilon$$

or

$$P(t > t_\epsilon) = \frac{1}{2}\epsilon \quad (14.5.4)$$

Remark. If the sample is large and σ unknown, we may also replace σ in (14.5.1) by its estimate s (or S) to give an *approximate* confidence interval

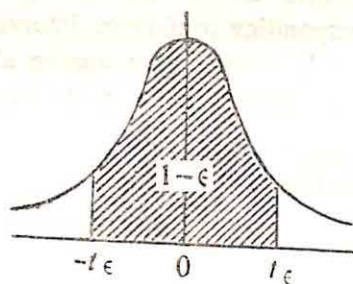


Fig. 26

$$\left(\bar{x} - \frac{su_\epsilon}{\sqrt{n}}, \bar{x} + \frac{su_\epsilon}{\sqrt{n}}\right) \quad (14.5.5)$$

This was, in fact, the usual practice in older statistics when Student's ratio was unknown. Now it can be theoretically proved that t -distributions with large degrees of freedom approximate to the standard normal distribution, and hence for large samples $t_\epsilon \simeq u_\epsilon$ so that the interval (14.5.5) is approximately the same as the exact result (14.5.3).

Example 2. Seven laboratory determinations of the value of g , the acceleration due to gravity at Calcutta gave a mean 977.51 cm/sec² and a standard deviation 4.42 cm/sec². Now it is known that the population of the measured values of any physical quantity subject to experimental errors has a normal distribution whose mean is the true value of the quantity (cf. Ch. 18, Theory of Errors). Assuming this fact, find 95% confidence interval for the true value of g .

For $\epsilon = .05$ and $\nu = 6$ degrees of freedom, by (14.5.4) and Table III, $t_\epsilon = 2.447$. Since $\bar{x} = 977.51$ and $S = 4.42$, a 95% confidence interval for the population mean is (973.09, 981.93).

Confidence interval for σ

The suitable statistic is

$$\chi^2 = \frac{nS^2}{\sigma^2}$$

whose sampling distribution has a χ^2 -distribution with $\nu = n - 1$ degrees of freedom.

Choose any positive number $\chi^2_{\epsilon_1}$, and determine $\chi^2_{\epsilon_2}$ by

$$P(\chi^2_{\epsilon_1} < \chi^2 < \chi^2_{\epsilon_2}) = 1 - \epsilon$$

This equation gives $\chi^2_{\epsilon_2}$ as a function of $\chi^2_{\epsilon_1}$. Or we have

$$P\left(\chi^2_{\epsilon_1} < \frac{nS^2}{\sigma^2} < \chi^2_{\epsilon_2}\right) = 1 - \epsilon$$

or

$$P\left(S\sqrt{\frac{n}{\chi^2_{\epsilon_2}}} < \sigma < S\sqrt{\frac{n}{\chi^2_{\epsilon_1}}}\right) = 1 - \epsilon$$

Hence

$$\left(S\sqrt{\frac{n}{\chi^2_{\epsilon_2}}}, S\sqrt{\frac{n}{\chi^2_{\epsilon_1}}}\right) \quad (14.5.6)$$

is a confidence interval for σ having confidence coefficient $1 - \epsilon$.

Now corresponding to different initial choices of $\chi^2_{\epsilon_1}$ we shall get different confidence intervals. Of these the shortest interval is obtained by minimising the length of the interval (14.5.6), viz.

$$\sqrt{nS} \left(\frac{1}{\sqrt{\chi^2_{\epsilon_1}}} - \frac{1}{\sqrt{\chi^2_{\epsilon_2}}} \right)$$

as a function of $\chi^2_{\epsilon_1}$. A practical determination of this shortest interval will be, however, very complicated, and is usually avoided in practice. It is instead customary to determine $\chi^2_{\epsilon_1}$, $\chi^2_{\epsilon_2}$ from the simple equations

$$P(0 < \chi^2 < \chi^2_{\epsilon_1}) = \frac{1}{2}\epsilon$$

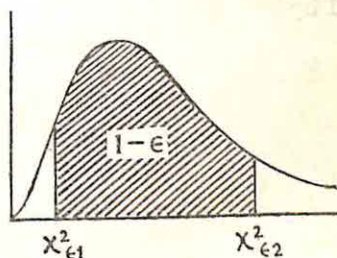


Fig. 27

or

$$P\left(\frac{X}{n} - u_\epsilon \sqrt{\frac{v(n-v)}{n^3}} < p < \frac{X}{n} + u_\epsilon \sqrt{\frac{v(n-v)}{n^3}}\right) \simeq 1 - \epsilon$$

which gives the approximate confidence interval (14.6.3).

2. Confidence interval for the mean of any population for large samples. Let us here digress a little from parametric estimation to consider the problem of interval estimation of a population characteristic, say, the mean for large samples. We know that for large samples the sample mean \bar{X} is approximately normal $(m, \sigma/\sqrt{n})$, and replacing σ by its estimate s (or S) \bar{X} can be taken to be approximately normal $(m, s/\sqrt{n})$, or $\frac{\sqrt{n}(\bar{X} - m)}{s}$ to be approximately normal $(0, 1)$. Hence

$$P\left(-u_\epsilon < \frac{\sqrt{n}(\bar{X} - m)}{s} < u_\epsilon\right) \simeq 1 - \epsilon$$

where u_ϵ is given by (14.6.1). This leads to the approximate confidence interval

$$\left(\bar{x} - \frac{su_\epsilon}{\sqrt{n}}, \bar{x} + \frac{su_\epsilon}{\sqrt{n}}\right) \quad (14.6.4)$$

for the population mean m .

This method is, of course, not restricted to the mean only, but may be used for other population characteristics as well.

14.7 EXERCISES

1. Estimate the parameter a of a continuous population having the density function $(1+a)x^a$ ($0 < x < 1$) by the method of maximum likelihood.

2. Prove that the maximum likelihood estimate of the parameter a of a population having density function $2(a-x)/a^2$ ($0 < x < a$) for a sample of unit size is $2x$, x being the sample value, and show that the estimate is biased.

3. A random variable X can take all non-negative integral values, and

$$P(X=i) = p(1-p)^i \quad (i=0, 1, 2, \dots)$$

where p ($0 < p < 1$) is a parameter. Find the maximum likelihood estimate of p on the basis of a sample of size n from the population of X .

4. Estimate the parameter μ of the Pascal distribution (Ex. 8 Sec. 7.14) by the method of maximum likelihood, and show that the estimate is unbiased and consistent.

5. Find the maximum likelihood estimate of σ^2 for a normal (m, σ) population if m is known, and show that the estimate is unbiased and consistent.

6. Considering a sample of unit size from the population of the sample mean for a normal (m, σ) population, find the maximum likelihood estimate of m , assuming σ to be known.

7. Let $s_1^2, s_2^2, \dots, s_k^2$ denote a sample of size k from the population of the unbiased estimate of the population variance for a normal population. Write down the likelihood function for this sample, and show that the maximum likelihood estimate of σ^2 is given by

$$\hat{\sigma}^2 = (s_1^2 + s_2^2 + \dots + s_k^2)/k$$

8. A population is defined by the density function

$$f(x; a) = \frac{x^{l-1} e^{-x/a}}{\Gamma(l) a^l} \quad (0 < x < \infty)$$

l being a known constant. Estimate the parameter a by the method of maximum likelihood, and show that the estimate is consistent and unbiased.

9. Show that approximate confidence limits for large samples for the parameter μ of a Poisson population having confidence coefficient $1 - \epsilon$ are given by the roots of the quadratic equation in μ :

$$n(\bar{x} - \mu)^2 = u_\epsilon^2 \mu$$

which, to the order of $1/\sqrt{n}$, are approximately

$$\bar{x} \pm u_\epsilon \sqrt{\bar{x}/n}$$

where u_ϵ is given by (14.6.1)

10. Show that

$$\frac{\bar{x}}{l} \left(1 \pm \frac{u_\epsilon}{\sqrt{nl}} \right)$$

are approximate large-sample confidence limits for the parameter a of the population defined in Ex. 8, u_ϵ being given by (14.6.1).

11. The population of scores of 10-year old children in a psychological performance (Dearnborn Formboard) test is known to have a standard deviation 5.2. If a random sample of size 20 shows a mean of 16.9, find 95% confidence limits for the mean score of the population, assuming that the population is normal.

12. The marks obtained by 17 candidates in an examination have a mean 57 and variance 64. Find 99% confidence limits for the mean of the population of marks, assuming it to be normal.

13. The heights in inches of 8 students of a college, chosen at random, were as follows: 62.2, 62.4, 63.1, 63.2, 65.5, 66.2, 66.3, 66.5. Compute 95% and 98% confidence intervals for the mean and standard deviation of the population of

or

and

$$\left. \begin{aligned} P(\chi^2 > \chi^2_{\epsilon_1}) &= 1 - \frac{1}{2}\epsilon \\ P(\chi^2 > \chi^2_{\epsilon_2}) &= \frac{1}{2}\epsilon \end{aligned} \right\} \quad (14.5.7)$$

which state that the areas of the tails of the χ^2 -density curve on the left of $\chi^2_{\epsilon_1}$ and the right of $\chi^2_{\epsilon_2}$ are equal, each being equal to $\frac{1}{2}\epsilon$.

Example 3. In Ex. 2 find 95% confidence interval for the population standard deviation which, according to the theory of errors, gives an inverse measure of precision of the measuring process.

Here $\epsilon = .05$, $n = 7$, $S = 4.42$; $\chi^2_{\epsilon_1}$ and $\chi^2_{\epsilon_2}$ are given by

$$P(\chi^2 > \chi^2_{\epsilon_1}) = .975, \quad P(\chi^2 > \chi^2_{\epsilon_2}) = .025$$

corresponding to $\nu = 6$ degrees of freedom. From Table II, $\chi^2_{\epsilon_1} = 1.218$, $\chi^2_{\epsilon_2} = 14.626$, and a 95% confidence interval for σ , the population standard deviation is (3.06, 10.60).

14.6 APPROXIMATE CONFIDENCE INTERVALS

1. Binomial (n, p) population. Let x be an observed value of the parent random variable X , i.e. a sample of unit size from the corresponding population, and we propose to find an approximate confidence interval for p , assuming that n (known) is large.

By DeMoivre-Laplace theorem, for large n the distribution of the variate

$$\frac{X - np}{\sqrt{np(1-p)}}$$

is approximately normal (0, 1). Hence if the points $\pm u_\epsilon$ are determined by

$$\int_{-u_\epsilon}^{u_\epsilon} \phi(x) dx = 1 - \epsilon$$

or

$$\int_{u_\epsilon}^{\infty} \phi(x) dx = \frac{1}{2}\epsilon \quad (14.6.1)$$

where $\phi(x)$ is the standard normal density function, then

$$P\left(-u_\epsilon < \frac{X-np}{\sqrt{np(1-p)}} < u_\epsilon\right) \simeq \int_{-u_\epsilon}^{u_\epsilon} \phi(x)dx = 1 - \epsilon$$

This can be re-set in the form

$$P(A < p < B) \simeq 1 - \epsilon$$

where A, B are the roots of the quadratic equation in p :

$$(X-np)^2 = u_\epsilon^2 np(1-p)$$

so that an approximate confidence interval for p is (a, b) where a, b are the roots of the equation

$$(v-np)^2 = u_\epsilon^2 np(1-p)$$

or

$$a, b = \frac{n(2v + u_\epsilon^2) \pm u_\epsilon \sqrt{4nv(n-v) + n^2 u_\epsilon^2}}{2n(n + u_\epsilon^2)} \quad (14.6.2)$$

Now if we calculate only up to the order of $1/\sqrt{n}$,

$$a, b \simeq \frac{v}{n} \pm u_\epsilon \sqrt{\frac{v(n-v)}{n^3}}$$

i.e. an approximate confidence interval for p is

$$\left(\frac{v}{n} - u_\epsilon \sqrt{\frac{v(n-v)}{n^3}}, \frac{v}{n} + u_\epsilon \sqrt{\frac{v(n-v)}{n^3}} \right) \quad (14.6.3)$$

Another method. The result (14.6.3) can be deduced more easily from a slightly different point of view. If n is large, we know that X is approximately normal $(np, \sqrt{np(1-p)})$. Now the standard deviation contains the parameter p which if replaced by its estimate $\hat{p} = v/n$ gives an approximate value of the standard deviation to be $\sqrt{v(n-v)/n}$. Thus in this two-fold approximation, one can take X to be approximately normal $(np, \sqrt{v(n-v)/n})$ for large values of n . Hence the variate

$$\frac{X-np}{\sqrt{v(n-v)/n}}$$

is approximately standard normal, and if u_ϵ is given by (14.6.1), we have

$$P\left(-u_\epsilon < \frac{X-np}{\sqrt{v(n-v)/n}} < u_\epsilon\right) \simeq 1 - \epsilon$$

heights of the students of the college, assuming it to be normal, and find the length of the interval in each case.

14. 171 out of 300 voters picked at random from a large electorate said that they were going to vote a particular candidate. Find 95% confidence interval for the proportion of voters of the electorate who would vote in favour of the candidate.

15. In Ex. 3 Sec. 12.6 find 99% confidence limits for the probability of obtaining 'six' in a throw with the die.

16. A sample of size 500 from a Poisson- μ population has a mean of 4.78. Calculate 95% confidence limits for μ .

17. In Ex. 4 Sec. 12.6 find a 95% confidence interval for the mean number of daily telephone calls, (a) assuming that the corresponding population has a Poisson distribution and (b) without assuming anything regarding the population distribution.

18. The weekly wages of 144 workers of a large factory were recorded, and the sample mean and standard deviation were found to be Rs. 23.52 and Rs. 6.71 respectively. Find 95% confidence limits for the mean wage. (Do not assume that the population of wages is normal.)

19. From two normal populations having parameters (m_1, σ) and (m_2, σ) , two independent samples of sizes n_1 and n_2 are respectively drawn. Find confidence limits for $m_1 - m_2$ having confidence coefficient $1 - \epsilon$. (Use the theorem of Sec. 16.7.)

20. Let v_1 and v_2 be independent samples of unit size from two binomial populations (n_1, p_1) and (n_2, p_2) respectively, where n_1, n_2 are known and large. Prove that confidence limits for $p_1 - p_2$ having confidence coefficient $1 - \epsilon$ are approximately

$$v_1/n_1 - v_2/n_2 \pm u_\epsilon \sqrt{v_1(n_1 - v_1)/n_1^3 + v_2(n_2 - v_2)/n_2^3}$$

where u_ϵ is given by (14.6.1).

BIVARIATE SAMPLES

15.1 SAMPLE FROM A BIVARIATE POPULATION

Let X and Y be a pair of random variables defined on the event space of a random experiment E . Any performance of E will give an observed value of the two-dimensional variable (X, Y) , and corresponding to n independent repetitions of E we get n observed values

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

of (X, Y) , which is a sample of size n from the bivariate population. The sample being random, the above sample values may be regarded as observed values of the n two-dimensional random variables

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$$

respectively which are mutually independent all having the same

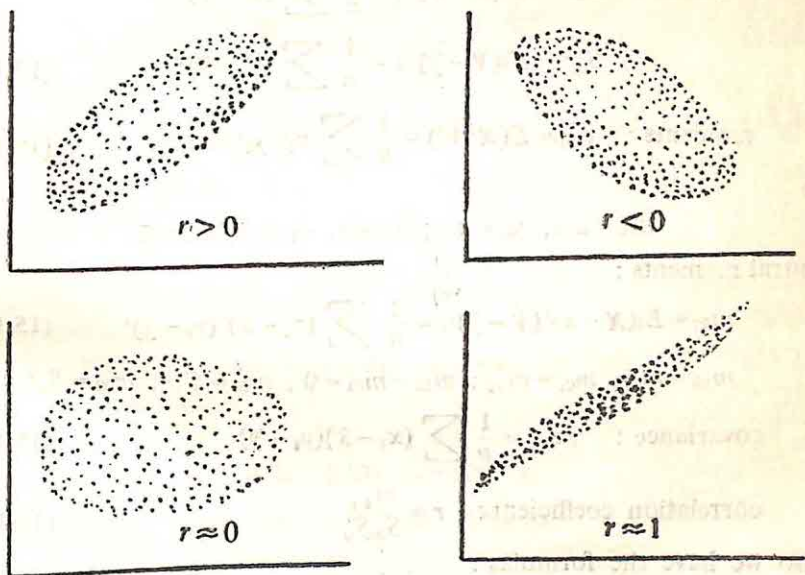


Fig. 28. Dot Diagrams for Different Values of r

distribution function, viz the distribution function of the population, $F(x, y)$.

By plotting the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ we obtain a diagram known as the *dot diagram* or *scatter diagram* of the sample. The scatter diagram usually looks like a cluster of dots on the xy -plane, particularly if the sample is from a continuous population and provides a simple but useful graphical representation of the data. We may also construct the three-dimensional analogues of the frequency diagram, histogram etc. but these are rather unwieldy.

The empirical distribution of the sample is obtained by placing a probability mass $1/n$ at each observed point (x_i, y_i) ($i = 1, 2, \dots, n$). Let (\dot{X}, \dot{Y}) denote the hypothetical random variable associated with the empirical distribution. The characteristics of (\dot{X}, \dot{Y}) are then the sample characteristics by definition, the most important of which are the following.

$$\text{means : } \bar{x} = E(\dot{X}) = \frac{1}{n} \sum x_i, \quad \bar{y} = E(\dot{Y}) = \frac{1}{n} \sum y_i \quad (15.1.1)$$

$$\begin{aligned} \text{variances : } S_x^2 &= E\{(\dot{X} - \bar{x})^2\} = \frac{1}{n} \sum (x_i - \bar{x})^2 \\ S_y^2 &= E\{(\dot{Y} - \bar{y})^2\} = \frac{1}{n} \sum (y_i - \bar{y})^2 \end{aligned} \quad (15.1.2)$$

$$\text{moments : } a_{kl} = E(\dot{X}^k \dot{Y}^l) = \frac{1}{n} \sum x_i^k y_i^l \quad (15.1.3)$$

So

$$a_{k0} = a_{xk}, \quad a_{0l} = a_{yl}; \quad a_{00} = 1, \quad a_{10} = \bar{x}, \quad a_{01} = \bar{y}$$

central moments :

$$m_{kl} = E\{(\dot{X} - \bar{x})^k (\dot{Y} - \bar{y})^l\} = \frac{1}{n} \sum (x_i - \bar{x})^k (y_i - \bar{y})^l \quad (15.1.4)$$

$$m_{k0} = m_{xk}, \quad m_{0l} = m_{yl}; \quad m_{10} = m_{01} = 0; \quad m_{20} = S_x^2, \quad m_{02} = S_y^2$$

$$\text{covariance : } m_{11} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) \quad (15.1.5)$$

$$\text{correlation coefficient : } r = \frac{m_{11}}{S_x S_y} \quad (15.1.6)$$

Also we have the formulas :

$$S_x^2 = a_{x2} - \bar{x}^2, \quad S_y^2 = a_{y2} - \bar{y}^2, \quad m_{11} = a_{11} - \bar{x}\bar{y} \quad (15.1.7)$$

15.2 PRACTICAL COMPUTATION

Discrete population. When the population is discrete, the sets of values x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n generally contain many repetitions. Let the distinct x - and y -values be ξ_j ($j = 1, 2, \dots, r$) and η_k ($k = 1, 2, \dots, s$), and let v_{jk} denote the frequency of occurrence of (ξ_j, η_k) in the sample. (If, however, a particular combination of ξ_j with η_k does not occur, the corresponding frequency is taken to be zero.) The data can now be arranged in a two-way frequency table showing v_{jk} columnwise against ξ_j and rowwise against η_k . Such a table is sometimes called a *correlation table* in view of the fact that we are mostly interested in computing the sample correlation coefficient r from this table.

If v_{xj} denotes the frequency of ξ_j in the set x_1, x_2, \dots, x_n and v_{yk} that of η_k in the set y_1, y_2, \dots, y_n , then

$$v_{xj} = \sum_k v_{jk}, \quad v_{yk} = \sum_j v_{jk}$$

so that

$$\sum v_{xj} = \sum v_{yk} = \sum \sum v_{jk} = n$$

Hence

$$\bar{x} = \frac{1}{n} \sum v_{xj} \xi_j, \quad \bar{y} = \frac{1}{n} \sum v_{yk} \eta_k \quad (15.2.1)$$

$$a_{x2} = \frac{1}{n} \sum v_{xj} \xi_j^2, \quad a_{yk} = \frac{1}{n} \sum v_{yk} \eta_k^2 \quad (15.2.2)$$

and

$$a_{11} = \frac{1}{n} \sum \sum v_{jk} \xi_j \eta_k$$

Set

$$p_k = \sum_j v_{jk} \xi_j, \quad q_j = \sum_k v_{jk} \eta_k \quad (15.2.3)$$

so that

$$\sum p_k = \sum \sum v_{jk} \xi_j = \sum v_{xj} \xi_j$$

$$\sum q_j = \sum \sum v_{jk} \eta_k = \sum v_{yk} \eta_k$$

and

$$\sum p_k \eta_k = \sum q_j \xi_j = \sum \sum v_{jk} \xi_j \eta_k$$

Then

$$a_{11} = \frac{1}{n} \sum p_k \eta_k = \frac{1}{n} \sum q_j \xi_j \quad (15.2.4)$$

To the given table we add columns for $v_y, v_y \eta, v_y \eta^2, p, p\eta$ and rows for $v_x, v_x \xi, v_x \xi^2, q, q\xi$ and obtain the respective totals. The following identities may be used as checks on the computation.

CHECK FORMULAS

$$\begin{aligned} (i) \quad \Sigma v_{xj} &= \Sigma v_{yk} & (ii) \quad \Sigma p_k &= \Sigma v_{xj} \xi_j \\ (iii) \quad \Sigma q_j &= \Sigma v_{yk} \eta_k & (iv) \quad \Sigma p_k \eta_k &= \Sigma q_j \xi_j \end{aligned} \quad (15.2.5)$$

$\bar{x}, \bar{y}, a_{x2}, a_{y2}$ and a_{11} are easily calculated from the table; S_x, S_y and m_{11} are then given by (15.1.7)

If convenient, we can make linear transformations

$$x_i = ax_i' + b, \quad y_i = cy_i' + d \quad (15.2.6)$$

i.e.

$$\xi_j = a\xi_j' + b, \quad \eta_k = c\eta_k' + d$$

for which the transformation formulas are

$$\bar{x} = a\bar{x}' + b, \quad \bar{y} = c\bar{y}' + d \quad (15.2.7)$$

$$S_x = |a| S_x', \quad S_y = |c| S_y', \quad r = r' \quad (\text{if } ac > 0)$$

Continuous population. A sample from a continuous population usually consists of more or less distinct values and if the size of the sample is not very large, we can draw up a two-column xy -table. For calculating the above characteristics columns for x^2, y^2 and xy only have be added. We may also make linear transformations if found suitable.

If, however, the sample is large, then grouping is necessary. Both the x - and y -values are grouped into classes and the results exhibited in a grouped correlation table. The sample characteristics can then be approximately calculated by treating the class mid-points for the x - and y -values as ξ 's and η 's respectively of the discrete case.

Example. The following correlation table shows the heights (in.) and weights (lb) of 114 adult males. Calculate the sample correlation coefficient between height and weight.

	Height (in.)									
Weight (lb)	56½- 58	58½- 60	60½- 62	62½- 64	64½- 66	66½- 68	68½- 70	70½- 72	72½- 74	Total
80½- 90			1							1
90½-100	1		2	2						5
100½-110			1	4	4	1	1			11
110½-120		1	3	4	8	4	1	1		22
120½-130			1	3	7	9	3		1	24
130½-140				1	9	4	4	1		19
140½-150				2	3	5	3	2	2	17
150½-160					1	2	1		1	5
160½-170					1	1	1	2	1	6
170½-180							1			1
180½-190						1	1			2
190½-200					1					1
Total	1	1	8	16	34	27	16	6	5	114

The tabular computation is shown in the next page.

By (15.2.1), (15.2.2) and (15.2.4)

$$\begin{aligned}\bar{x}' &= 0.508772, & \bar{y}' &= -0.535088 \\ a_{x_2}' &= 2.614035, & a_{y_2}' &= 4.464912 \\ a_{11}' &= 1.350877\end{aligned}$$

By (15.1.7)

$$\begin{aligned}S_{x'}^2 &= 2.355186 \quad \text{or} \quad S_{x'} = 1.534661 \\ S_{y'}^2 &= 4.178593 \quad \text{or} \quad S_{y'} = 2.044161 \\ m_{11}' &= 1.623115\end{aligned}$$

which give

$$r' = 0.517394$$

Hence the correlation coefficient together with other characteristics of the sample are given according to (15.2.7) by

$$\begin{aligned}\bar{x} &= 66.080044, & \bar{y} &= 129.89912 \\ S_x &= 3.069322, & S_y &= 20.44161 \\ r &= 0.517394\end{aligned}$$

or

$$\bar{x} = 66.08, \quad \bar{y} = 129.90, \quad S_x = 3.07, \quad S_y = 20.44, \quad r = 0.517$$

15.3 LEAST SQUARE CURVE FITTING

For fitting a curve of the type

$$y = g(x; c_0, c_1, \dots)$$

where c_0, c_1, \dots are unknown parameters, to the empirical sample distribution, i.e. to the observed set of points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ by the principle of least squares, we have to minimise

$$E[\{\hat{Y} - g(\hat{X}; c_0, c_1, \dots)\}^2] = \frac{1}{n} \sum \{y_i - g(x_i; c_0, c_1, \dots)\}^2$$

or to minimise $\sum \{y_i - g(x_i; c_0, c_1, \dots)\}^2$ as a function of c_0, c_1, \dots . This is the empirical form of the principle of least squares which consists in making the sum of the squares of the deviations of the observed points from the curve measured in the direction of the y -axis a minimum.

Regression lines. It follows from the general theory (Sec. 8.13) that the regression lines of the sample are

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad (15.3.1)$$

and

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

the former being the regression line of \hat{Y} on \hat{X} and the latter that of \hat{X} on \hat{Y} , where

$$b_{yx} = r \frac{S_y}{S_x}, \quad b_{xy} = r \frac{S_x}{S_y} \quad (15.3.2)$$

are the *regression coefficients* of the sample.

It also follows that $|r|$ gives a measure of goodness of fit of the regression lines to the observed points.

Example 1. Find the regression lines of the sample in the example of Sec. 15.2.

Here

$$b_{yx} = 3.44583, \quad b_{xy} = 0.07768702$$

and hence the regression lines are

$$y - 129.90 = 3.45(x - 66.80)$$

$$x - 66.08 = 0.0777(y - 129.90)$$

Parabolic curve fitting. The normal equations for fitting a k th degree parabola

$$y = c_0 + c_1x + c_2x^2 + \dots + c_kx^k \quad (15.3.3)$$

to the observed points are, by (8.14.4)

$$\begin{aligned} c_0^*a_{00} + c_1^*a_{10} + c_2^*a_{20} + \dots + c_k^*a_{k0} &= a_{01} \\ c_0^*a_{10} + c_1^*a_{20} + c_2^*a_{30} + \dots + c_k^*a_{k+1,0} &= a_{11} \\ &\dots \quad \dots \quad \dots \end{aligned} \quad (15.3.4)$$

$$c_0^*a_{k0} + c_1^*a_{k+1,0} + c_2^*a_{k+2,0} + \dots + c_k^*a_{2k,0} = a_{k1}$$

Or multiplying all the equations by n , we obtain

$$\begin{aligned} nc_0^* + c_1^*\sum x_i + c_2^*\sum x_i^2 + \dots + c_k^*\sum x_i^k &= \sum y_i \\ c_0^*\sum x_i + c_1^*\sum x_i^2 + c_2^*\sum x_i^3 + \dots + c_k^*\sum x_i^{k+1} &= \sum x_i y_i \\ &\dots \quad \dots \quad \dots \quad \dots \end{aligned} \quad (15.3.5)$$

$$c_0^*\sum x_i^k + c_1^*\sum x_i^{k+1} + c_2^*\sum x_i^{k+2} + \dots + c_k^*\sum x_i^{2k} = \sum x_i^k y_i$$

These equations determine the least square estimates $c_0^*, c_1^*, \dots, c_k^*$ of the parameters.

The *residual* of \bar{Y} , \bar{V}_y is given by

$$\bar{V}_y = \bar{Y} - c_0^* - c_1^*\bar{X} - \dots - c_k^*\bar{X}^k \quad (15.3.6)$$

and the *residuals of the sample*, i.e. the values v_{yi} which \bar{V}_y takes are given by

$$v_{yi} = y_i - c_0^* - c_1^*x_i - \dots - c_k^*x_i^k \quad (i = 1, 2, \dots, n) \quad (15.3.7)$$

1. The first equation of (15.3.5) states that $\sum v_{yi} = 0$, i.e. the sum of the residuals is zero.

2. The normal equations may also be written as

$$\sum v_{yi} = 0, \sum x_i v_{yi} = 0, \dots, \sum x_i^k v_{yi} = 0 \quad (15.3.8)$$

Assuming, for simplicity of discussions, that all the x_i 's are distinct, the normal equations will yield determinate solutions only if $k \leq n-1$. For if $k \geq n-1$ or $n \leq k+1$, any n equations of the set (15.3.8) reduce to $v_{yi} = 0$ or

$$y_i = c_0^* + c_1^*x_i + \dots + c_k^*x_i^k \quad (i = 1, 2, \dots, n) \quad (15.3.9)$$

so that the rest of the $(k-n+1)$ equations are identically satisfied, which shows that only n of the $(k+1)$ normal equations are independent, and if $k > n-1$ or $n < k+1$, the n equations (15.3.9) in the $(k+1)$ unknowns c_0^* , c_1^* , ..., c_k^* are necessarily indeterminate. Geometrically, the equations (15.3.9) mean that the best-fitting k th degree parabola would pass through all the n observed points, and if $k > n-1$, an infinite number of such parabolas are obviously possible. If, however, $k = n-1$, the solutions of (15.3.9) are unique and the best-fitting parabola exactly passes through all the observed points, i.e. we may say that the case of least square fitting reduces to that of interpolation.

A suitable measure of goodness of fit of the best-fitting parabola to the observed points is provided by

$$\left. \begin{array}{l} \text{where} \\ R_y = \rho(\hat{U}_y, \hat{Y}) \\ \hat{U}_y = c_0^* + c_1^* \hat{X} + \dots + c_k^* \hat{X}^k \end{array} \right\} \quad (15.3.10)$$

From (8.14.12) and (8.14.13), $0 \leq R_y \leq 1$ and

$$R_y^2 = 1 - \frac{\sum v_{yi}^2}{nS_y^2} \quad (15.3.11)$$

where, by (8.14.10)

$$\sum v_{yi}^2 = \sum y_i^2 - c_0^* \sum y_i - c_1^* \sum x_i y_i - \dots - c_k^* \sum x_i^k y_i \quad (15.3.12)$$

Practical computation. Let us suppose, for convenience, that the data is presented in a two-column xy -table. For the normal equations (15.3.5), we shall have to prepare columns for x^2 , x^3 , ..., x^{2k} ; xy , x^2y , ..., $x^k y$ and a further column for y^2 for computing R_y .

Linear transformations $x = ax' + b$, $y = cy' + d$ are sometimes helpful. In that case we first find the best-fitting parabola to the points (x'_i, y'_i) ($i = 1, 2, \dots, n$), which when transformed back in terms of x, y gives the best-fitting parabola to the original data. We note that this depends upon the fact that by the above transformations a general parabola in the (x, y) -plane is transformed again into a general parabola in the (x', y') -plane and vice versa. It can be easily seen that the measure of goodness of fit R_y remains invariant under the above transformations.

Example 2. The percentages of protein content (x) and vitreous kernel (y) in 8 samples of wheat were found to be as follows :

x	24	36	45	55	75	84	91	96
y	10.1	10.2	10.8	11.1	12.2	13.3	14.0	15.8

Fit a parabola of the form $y = c_0 + c_1x + c_2x^2$ to the above data.

$$\text{Set } y = y' + 10.$$

x	y	y'	x^2	x^3	x^4	xy'	x^2y'	y'^2
24	10.1	0.1	576	13824	331776	2.4	57.6	0.01
36	10.2	0.2	1296	46656	1679616	7.2	259.2	0.04
45	10.8	0.8	2025	91125	4100625	36.0	1620.0	0.64
55	11.1	1.1	3025	166375	9150625	60.5	3327.5	1.21
75	12.2	2.2	5625	421875	31640625	165.0	12375.0	4.84
84	13.3	3.3	7056	592704	49787136	277.2	23284.8	10.89
91	14.0	4.0	8281	753571	68574961	364.4	33124.0	16.00
96	15.8	5.8	9216	884736	84934656	556.8	53452.8	33.64
506	—	17.5	37100	2970866	250200020	1469.1	127500.9	67.27

The normal equations are

$$8c_0' + 506c_1' + 37100c_2' = 17.5$$

$$506c_0' + 37100c_1' + 2970866c_2' = 1469.1$$

$$37100c_0' + 2970866c_1' + 250200020c_2' = 127500.9$$

Solving these we get

$$c_0' = 1.371707, c_1' = -0.07382217, c_2' = 0.001182759$$

Hence the equation of the parabola is

$$y' = 1.372 - 0.07382x + 0.001183x^2$$

or

$$y = 11.372 - 0.07382x + 0.001183x^2$$

which is the required best-fitting parabola.

Now by (15.3.12) $\Sigma v_{y,i}^2 = 0.9144$ and

$$\bar{y}' = 2.187500, a_{y^2}' = 8.408750, S_{y'}^2 = 3.623594$$

Hence by (15.3.11)

$$R_{y'} = 0.984 \text{ or } R_y = 0.984$$

This shows that the fit of the above parabola to the observed points is quite good.

The correlation coefficient of the sample is found to be 0.942 which shows that the regression lines also fit well to the data, but the parabolic fit is much better.

15.4 MAXIMUM LIKELIHOOD ESTIMATION

The logic of the process is exactly the same as in the univariate case. Here the likelihood function of the sample $L = L(x_1, \dots, x_n, y_1, \dots, y_n; \theta_1, \dots, \theta_k)$, $\theta_1, \dots, \theta_k$ being the population parameters, is defined as follows.

DISCRETE CASE. $L = f_{x_1, y_1}(\theta_1, \dots, \theta_k) \dots \dots \dots f_{x_n, y_n}(\theta_1, \dots, \theta_k)$ where $f_{x_i, y_i}(\theta_1, \dots, \theta_k) = P(X = x_i, Y = y_i)$.

CONTINUOUS CASE. $L = f(x_1, y_1; \theta_1, \dots, \theta_k) \dots \dots \dots f(x_n, y_n; \theta_1, \dots, \theta_k)$ where $f(x, y; \theta_1, \dots, \theta_k)$ is the density function of (X, Y) .

For fixed sample values we maximise L or $\log L$, the equations for which are

$$\frac{\partial \log L}{\partial \theta_1} = 0, \dots, \frac{\partial \log L}{\partial \theta_k} = 0$$

A set of solutions of these likelihood equations :

$$\theta_1 = \hat{\theta}_1(x_1, \dots, x_n, y_1, \dots, y_n), \dots, \theta_k = \hat{\theta}_k(x_1, \dots, x_n, y_1, \dots, y_n)$$

which corresponds to a unique maximum of L then gives the required maximum likelihood estimates of the parameters.

Bivariate normal population. The density function of the population is given by (6.4.3), and hence

$$L = (2\pi)^{-n} \sigma_x^{-n} \sigma_y^{-n} (1 - \rho^2)^{-n/2} \times e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{(x_i - m_x)^2}{\sigma_x^2} - 2\rho \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y} + \frac{(y_i - m_y)^2}{\sigma_y^2} \right\}} \quad (15.4.1)$$

or

$$\log L = -n \log (2\pi) - n \log \sigma_x - n \log \sigma_y - \frac{1}{2} n \log (1 - \rho^2) - \frac{1}{2(1-\rho^2)} \sum \left\{ \frac{(x_i - m_x)^2}{\sigma_x^2} - 2\rho \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y} + \frac{(y_i - m_y)^2}{\sigma_y^2} \right\}$$

The likelihood equations are

$$\begin{aligned} (i) \quad \frac{\partial \log L}{\partial m_x} &= 0, & (ii) \quad \frac{\partial \log L}{\partial m_y} &= 0, & (iii) \quad \frac{\partial \log L}{\partial \sigma_x} &= 0 \\ (iv) \quad \frac{\partial \log L}{\partial \sigma_y} &= 0, & (v) \quad \frac{\partial \log L}{\partial \rho} &= 0 \end{aligned}$$

(i) and (ii) give

$$-\frac{\sum(x_i - m_x)}{\sigma_x} + \rho \frac{\sum(y_i - m_y)}{\sigma_y} = 0$$

$$\frac{\sum(x_i - m_x)}{\sigma_x} - \frac{\sum(y_i - m_y)}{\sigma_y} = 0$$

Since $\rho^2 \neq 1$, we must have

$$\sum(x_i - m_x) = 0, \quad \sum(y_i - m_y) = 0$$

or

$$\hat{m}_x = \bar{x}, \quad \hat{m}_y = \bar{y} \quad (vi)$$

(iii) and (iv) respectively reduce to

$$n + \frac{1}{1 - \rho^2} \sum \left\{ -\frac{(x_i - m_x)^2}{\sigma_x^2} + \rho \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y} \right\} = 0$$

and

$$n + \frac{1}{1 - \rho^2} \sum \left\{ \rho \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y} - \frac{(y_i - m_y)^2}{\sigma_y^2} \right\} = 0$$

or using (vi) we get

$$1 - \rho^2 - \frac{S_x^2}{\sigma_x^2} + \rho r \frac{S_x S_y}{\sigma_x \sigma_y} = 0$$

and

$$1 - \rho^2 - \frac{S_y^2}{\sigma_y^2} + \rho r \frac{S_x S_y}{\sigma_x \sigma_y} = 0$$

So

$$\frac{S_x^2}{\sigma_x^2} = \frac{S_y^2}{\sigma_y^2} = \frac{S_x S_y}{\sigma_x \sigma_y} = k \text{ (say)} \quad (vii)$$

and

$$1 - \rho^2 = k(1 - \rho r) \quad (viii)$$

(v) gives

$$\begin{aligned} n\rho - \frac{\rho}{1 - \rho^2} \sum \left\{ \frac{(x_i - m_x)^2}{\sigma_x^2} - 2\rho \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y} + \frac{(y_i - m_y)^2}{\sigma_y^2} \right\} \\ + \frac{1}{\sigma_x \sigma_y} \sum (x_i - m_x)(y_i - m_y) = 0 \end{aligned}$$

Using (vi) this reduces to

$$\rho(1 - \rho^2) - \rho \left(\frac{S_x^2}{\sigma_x^2} - 2\rho r \frac{S_x S_y}{\sigma_x \sigma_y} + \frac{S_y^2}{\sigma_y^2} \right) + (1 - \rho^2)r \frac{S_x S_y}{\sigma_x \sigma_y} = 0$$

or by (vii)

$$(1 - \rho^2)(\rho + kr) = 2\rho k(1 - \rho r)$$

This together with (viii) gives $\rho = kr$ which when substituted again in (viii) shows that $k = 1$, and hence

$$\hat{m}_x = \bar{x}, \hat{m}_y = \bar{y}, \hat{\sigma}_x = S_x, \hat{\sigma}_y = S_y, \hat{\rho} = r$$

15.5 EXERCISES

1. The following table gives the evaporation values in mm. from two evaporimeter tanks, one of which is mesh-covered and the other kept in a cage. Find the correlation coefficient and the regression lines of the sample.

Cage value (mm.)	Mesh-covered value (mm.)									Total
	3.5 -4.5	4.5 -5.5	5.5 -6.5	6.5 -7.5	7.5 -8.5	8.5 -9.5	9.5 -10.5	10.5 -11.5	11.5 -12.5	
3.5 - 4.5	2	1								3
4.5 - 5.5		1	1	1						3
5.5 - 6.5			2	2	1					5
6.5 - 7.5				3	3					6
7.5 - 8.5					2	4	5			11
8.5 - 9.5						5	5	2		12
9.5 - 10.5							9	10	3	22
10.5 - 11.5								3	6	9
Total	2	2	3	6	6	9	19	15	9	71

2. Compute the correlation coefficient between the marks (%) obtained in Numerical Analysis Theoretical and Practical at the M.A. and M.Sc. Examination in Applied Mathematics 1961 by 14 candidates given by the table in the next page, and find the lines of regression.

Theoretical mark	Practical mark	Theoretical mark	Practical mark
84	92	46	82
76	56	40	30
72	84	38	70
70	94	38	54
64	88	34	88
54	66	30	52
52	90	28	60
48	78	22	58

3. A new-born baby was weighed weekly from birth, and 9 such weights (y) in ounces against ages (x) in weeks are shown below.

x	0	1	2	3	4	5	6	7	8
y	119	141	144	149	150	158	161	166	170

Find the regression line and the regression parabola of the second degree of the sample of weight on age, and calculate the goodness of fit in each case.

4. Fit a straight line (a) $y = c_0 + c_1x$ and parabolas (b) $y = c_0 + c_1x + c_2x^2$ and (c) $y = c_0 + c_2x^2$ to the following data, and compare their goodness of fit.

x	3.5	8.4	16.8	23.9	27.1	28.8
y	4.4	9.2	20.6	31.1	35.0	37.7

5. Fit a curve of the form $y = x^2 + ax + b$ to the following data by the method of least squares.

x	2	3	4	5	6
y	7.2	3.9	3.0	4.4	6.3

TESTING OF HYPOTHESES I

16.1 STATISTICAL HYPOTHESES—SIMPLE AND COMPOSITE

As stated earlier, the distribution of the population is usually unknown, and our task is to derive some rough knowledge about the same from a sample drawn from the population. The problem in its direct form constitutes the problem of estimation which we have already discussed. Now the problem also sometimes presents itself in an indirect but an equally important form, viz. testing of some a priori knowledge about the population distribution obtained from theoretical considerations or otherwise on the basis of a sample. This will be the subject of the present study.

A *statistical hypothesis* is, in general, defined to be an assumption of any sort about the distribution function of the population, $F(x)$. In this chapter we shall assume that $F(x)$ has a known functional form which involves a number of unknown parameters $\theta_1, \theta_2, \dots, \theta_k$. Any one will reflect that this assumption is in itself a statistical hypothesis which we shall not, however, test for the present but take to be granted with certainty. A hypothesis then consists in any assumption regarding the parameters $\theta_1, \theta_2, \dots, \theta_k$; for example, some or all of the parameters take prescribed values or lie in prescribed intervals, or one or more given relations exist between them and so on.

The hypotheses may be classified into two types—simple and composite. A hypothesis of the form

$$H_0 : \theta_j = \theta_{0j} \quad (j = 1, 2, \dots, k) \quad (16.1.1)$$

where θ_{0j} 's are given numbers, i.e. a hypothesis which prescribes exact values to all the parameters is called a *simple* hypothesis. And a hypothesis which is not simple is a *composite* hypothesis.

The above notions may be conveniently and precisely stated by means of a geometric formalism. If we write

$$\Theta = (\theta_1, \theta_2, \dots, \theta_k) \quad (16.1.2)$$

then Θ represents a point in a k -dimensional space called the *parametric space* P_k , Θ being called the *parametric point*. It is to be noted that corresponding to all possible or admissible values of the parameters, the parametric point Θ may or may not run over the entire space P_k , e.g. for a normal (m, σ) population the parametric space P_2 consists of the two-dimensional (m, σ) plane, but since σ is necessarily positive, the parametric point is confined to the half-plane $\sigma > 0$.

A statistical hypothesis may now be defined to be any assumption of the form

$$H_0 : \Theta \in \omega \quad (16.1.3)$$

where ω is a given set of points in P_k .

If ω consists of a single point $\Theta_0 = (\theta_{01}, \theta_{02}, \dots, \theta_{0k})$, then $H_0 : \Theta = \Theta_0$ or $\theta_j = \theta_{0j}$ ($j = 1, 2, \dots, k$), i.e. H_0 is a simple hypothesis. If ω consists of more than one point, the hypothesis H_0 is composite.

Example. Consider a normal (m, σ) population.

(a) $H_0 : m = 2, \sigma = 0.1$ is a simple hypothesis. The point set ω consists of the single point $(2, 0.1)$ in the parametric plane P_2 .

(b) $H_0 : m = 2$. Since σ is unspecified, ω consists of the straight line $m = 2$, and hence H_0 is a composite hypothesis.

(c) $H_0 : 2 < m < 3$ is a composite hypothesis. Here ω is the infinite strip lying between the parallels $m = 2$ and $m = 3$.

(d) $H_0 : m = \sigma^2$ is also a composite hypothesis, ω consisting of the points on the parabola $m = \sigma^2$.

Null hypothesis. We shall very often make a hypothesis wishing it to be rejected by the test. Such a hypothesis is called a *null hypothesis*. This may seem somewhat strange at the first instant, but we shall presently see that not only from the theoretical standpoint but also in many important types of practical problems statistical hypotheses appear naturally as null hypotheses. Some such practical problems will be cited in the course of our discussions.

Alternative hypothesis. Sometimes it so happens that we know for certain that either $\Theta \in \omega$ or $\Theta \in \omega_1$ where ω and ω_1 are two disjoint point sets in P_k , and it remains for us to decide between the

two by means of a test. Now if we have a priori reasons to be more inclined to believe in the latter hypothesis, then we set up the null hypothesis

$$\left. \begin{array}{l} H_0 : \Theta \in \omega \\ \text{to be tested against the alternative hypothesis} \\ H_1 : \Theta \in \omega_1 \end{array} \right\} \quad (16.1.4)$$

hoping that the null hypothesis will be rejected by the test and thereby confirm our belief in the alternative. If, however, we do not have sufficient reasons for favouring one hypothesis to the other, then we may set up any one of these as the null hypothesis and the other as the alternative.

In the general theory, therefore, we shall consider a null hypothesis H_0 against an alternative H_1 as in (16.1.4). In case an alternative is not stated, it naturally means that the alternative is the negation of the null hypothesis, i.e.

$$H_1 : \Theta \in \bar{\omega}$$

where $\bar{\omega}$ is the complement of ω in P_k .

16.2 GENERAL FORM OF A TEST. BEST CRITICAL REGION

Let a sample of size $n : x_1, x_2, \dots, x_n$ be drawn from the population. On the evidence offered by this sample we shall have to decide whether to accept or reject the null hypothesis. The mathematical formulation of this evidence is known as a *test* of the hypothesis H_0 . Since, as we have remarked earlier, it is impossible to make a decision with perfect certainty, we must also have to state how much of confidence can be placed on such a decision, or, we may say, how much the test is significant. As such these tests are often called *tests of significance*. It is, however, customary and also more convenient to measure the significance of a test not in terms of its degree of reliability but in terms of the complementary quantity—the amount of risk of misjudgment taken in pronouncing the decision.

We shall, for simplicity, confine our general discussions to the case of a continuous population only. Let $f(x; \theta_1, \theta_2, \dots, \theta_k) = f(x; \Theta)$ denote the density function of the population. The density function

of the sample point x is then identical with the likelihood function $L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = L(x; \Theta)^*$ given by (14.1.3).

A test of the hypothesis H_0 , in its general form, consists in choosing a region W in the sample space R^n such that if the observed position of the sample point x falls in W , H_0 is rejected, and if it falls in \bar{W} , the part of R^n outside W , H_0 is accepted; W is called the *rejection region* or the *critical region* and \bar{W} the *acceptance region* of the test.

Two types of error. Now the sample point x is a random variable, and as such an observed position of x may be in the critical region as well as outside it. Accordingly, we are liable to make two types of error of decision detailed below.

TYPE I ERROR: If H_0 is true, but x falls in the critical region W when we reject H_0 .

TYPE II ERROR: If H_0 is false (and hence H_1 is true), but x falls in the acceptance region \bar{W} when we accept H_0 .

The probability of committing Type I error is clearly

$$P(x \in W | \Theta \in \omega) \quad (16.2.1)$$

and that of Type II error is

$$P(x \in \bar{W} | \Theta \in \omega_1) = 1 - P(x \in W | \Theta \in \omega_1) = 1 - \beta(W) \quad (16.2.2)$$

where

$$\beta(W) = P(x \in W | \Theta \in \omega_1) \quad (16.2.3)$$

represents the probability of rejecting H_0 when it is false and is called the *power of the critical region W with respect to the alternative hypothesis H_1* or simply the *power of the test*.

Best critical region. For constructing good tests we naturally try to reduce the probabilities of both types of error as much as possible. But it is generally found that if the probability of one type of error is decreased, the probability of the other automatically increases, and it becomes impossible to make the probabilities of both

*Here, for convenience, we have denoted the observed value (x_1, x_2, \dots, x_n) of the sample point x by the same symbol x .

the errors arbitrarily small simultaneously. Hence, in order to obtain a useful test, we must strike a balance between the two types of error, and for this it is necessary to formulate a principle which would tell us the best way of doing the same. Now many such principles may be thought of as possible, of which the most satisfactory has been found to be the following.

1. First fix the probability of committing the Type I error arbitrarily, i.e. given any number ε ($0 < \varepsilon < 1$) in advance, set

$$P(x \in W | \theta = \theta_0) = \varepsilon \quad (16.2.4)$$

the number ε being called the *significance level* of the test. The equation (16.2.4), in general, gives rise to a family of critical regions at the significance level ε .

2. Now in the above family of critical regions choose, if possible, that particular critical region which would minimise the probability of committing Type II error, i.e. maximise the power of the test. This critical region, if existent, will be called the *best critical region* and the corresponding test the *best test* or the *most powerful test* at the given significance level ε .

16.3 BEST CRITICAL REGION FOR SIMPLE HYPOTHESES

The best critical region for testing a simple hypothesis

$$\left. \begin{array}{l} H_0 : \theta = \theta_0 \\ \text{against a simple alternative} \\ H_1 : \theta = \theta_1 \end{array} \right\} \quad (16.3.1)$$

is given by the following theorem.

Neyman-Pearson theorem. The set of all points x in the sample space R^n satisfying the inequality

$$\frac{L(x; \theta_0)}{L(x; \theta_1)} < k \quad (16.3.2)$$

is the best critical region $W = W(k)$, where $k(> 0)$ is a constant which is determined (if possible) as a function of the given significance level ε by

$$P(x \in W | \theta = \theta_0) = \varepsilon \quad (16.3.3)$$

Proof. Let W' be any other critical region at the significance level ε , i.e.

$$P(x \in W' | \theta = \theta_0) = \varepsilon \quad (i)$$

Clearly, the theorem will be proved if we can show that the power of the region W is greater than that of W' , i.e.

$$\beta(W) > \beta(W')$$

Form (i) and (16.3.3) we get

$$\int_W L(x; \theta_0) dx = \int_{W'} L(x; \theta_0) dx \quad [dx = dx_1 dx_2 \dots dx_n]$$

or

$$\int_{W - WW'} L(x; \theta_0) dx = \int_{W' - WW'} L(x; \theta_0) dx \quad (ii)$$

Now in $W - WW'$

$$L(x; \theta_1) > \frac{1}{k} L(x; \theta_0)$$

and in $W' - WW'$

$$L(x; \theta_1) \leq \frac{1}{k} L(x; \theta_0)$$

So

$$\begin{aligned} \beta(W) &= \int_W L(x; \theta_1) dx = \int_{W - WW'} L(x; \theta_1) dx + \int_{WW'} L(x; \theta_1) dx \\ &> \frac{1}{k} \int_{W - WW'} L(x; \theta_0) dx + \int_{WW'} L(x; \theta_1) dx \end{aligned}$$

and

$$\begin{aligned} \beta(W') &= \int_{W'} L(x; \theta_1) dx = \int_{W' - WW'} L(x; \theta_1) dx + \int_{WW'} L(x; \theta_1) dx \\ &\leq \frac{1}{k} \int_{W' - WW'} L(x; \theta_0) dx + \int_{WW'} L(x; \theta_1) dx \end{aligned}$$

Hence by (ii) $\beta(W) > \beta(W')$, which proves the theorem.

Composite alternative. The best critical region W , in general, depends on θ_1 . But if the situation is such that for all values of θ_1 lying in a k -dimensional interval $\alpha < \theta_1 < \beta$ we get the same best critical region W , then W can evidently be regarded as the best critical region for $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \alpha < \theta < \beta$.

Working rule. Now the above theorem gives an n -dimensional critical region, n being the size of the sample, which is rather impracticable to work with. Accordingly, we deduce the following convenient working rule.

Rearrange the inequality (16.3.2) in any convenient form

$$z \in R \quad (16.3.4)$$

where $z = z(x)$ is a suitable statistic whose density function $f_z(z; \theta)$ can be easily obtained and R is a region of the z -axis, usually consisting of one or two intervals, determined by

$$P(Z \in R | \theta = \theta_0) = \int_R f_z(z; \theta_0) dx = \varepsilon \quad (16.3.5)$$

Thus the n -dimensional critical region W dwindles into two things, viz. a statistic z and the corresponding one-dimensional critical region R .

To sum up we proceed along the following steps :

1. Write down the likelihood function $L(x; \theta)$.

2. Write the inequality (16.3.2), and reset it in the form (16.3.4). This gives the statistic z and the form of the critical region R . This can usually be done in infinitely many ways ; of these, choose a form of (16.3.4) which gives the most convenient statistic.

3. Determine R by (16.3.5). This fixes the extent of R which is thereby completely determined.

4. Compute the value of the statistic z from the sample. If this value falls in R , reject H_0 ; otherwise accept it. A computed value of the statistic which falls in the critical region is customarily said to be *significant*.

5. In a practical problem the significance level is not usually decided upon in advance. In such cases we calculate the value of

z , and find the minimum significance level at which the critical region just includes this value of z . If this level is sufficiently small, we can reject H_0 , and if not, we have to accept H_0 . The question how small this level should be in order to be able to reject H_0 confidently is, of course, entirely relative and depends on the particular nature of the problem we are dealing with. The following convention is, however, often adopted in practice.

A computed value of z will be called (i) *not significant* if it falls outside the critical region of the 5% level, (ii) *simply significant* if it falls inside the critical region of the 5% level but outside that of the 1% level and (iii) *highly significant* if it falls inside the critical region of the 1% level.

16.4 APPLICATIONS TO NORMAL (m, σ) POPULATION

Test for m . We assume that σ is *known* and wish to test the hypothesis

$$H_0 : m = m_0$$

against an alternative

$$H_1 : m = m_1$$

where m_0 and m_1 are two given unequal numbers.

From Sec. 14.2

$$L(x; m) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m)^2}$$

So

$$\begin{aligned} \frac{L(x; m_0)}{L(x; m_1)} &= e^{-\frac{1}{2\sigma^2} [\sum (x_i - m_0)^2 - \sum (x_i - m_1)^2]} \\ &= e^{-\frac{n}{2\sigma^2} (m_1 - m_0)(\bar{x} - m_0 - m_1)} \end{aligned}$$

Hence the form of the best critical region is given by

$$e^{-\frac{n}{2\sigma^2} (m_1 - m_0)(\bar{x} - m_0 - m_1)} < k$$

k being a constant.

Now two cases arise according as $m_1 >$ or $< m_0$.

Case I. $m_1 > m_0$. In this case the above inequality reduces to the form $\bar{x} > k'$, k' being another constant, or, more conveniently, to the form

$$u = \frac{\sqrt{n}(\bar{x} - m_0)}{\sigma} > u_\epsilon \quad (16.4.1)$$

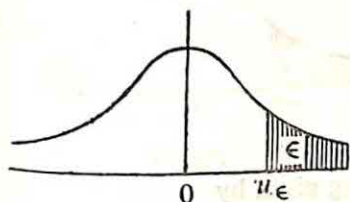


Fig. 29

where we know that, under H_0 , U has a standard normal distribution. Hence we may choose u as our required statistic, the corresponding best critical region being the interval (u_ϵ, ∞) where u_ϵ is determined by

$$P(U > u_\epsilon) = \epsilon \quad (16.4.2)$$

We note that the constant u_ϵ is directly determined by (16.4.2), and hence the statistic u and the corresponding critical region are independent of the particular value of m_1 but depend only on the prior assumption that $m_1 > m_0$. This obviously means that the above test serves as the best test of H_0 against the composite alternative $H_1 : m > m_0$.

Case II. $m_1 < m_0$. Here the inequality in question reduces to

$$u = \frac{\sqrt{n}(\bar{x} - m_0)}{\sigma} < -u_\epsilon \quad (16.4.3)$$

so that best critical region for u is $(-\infty, -u_\epsilon)$ where

$$P(U < -u_\epsilon) = \epsilon$$

or

$$P(U > u_\epsilon) = \epsilon$$

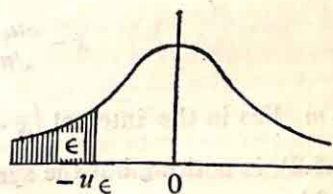


Fig. 30

which is the same as (16.4.2). This again serves as the best test of H_0 against the alternative $H_1 : m < m_0$.

In *Case I* the critical region is the right tail of the standard normal density curve, and in *Case II* it is the left tail of the same, and as such

we say that we are concerned with a right-tailed test in the former and a left-tailed test in the latter.

For testing $H_0 : m = m_0$ against no specific alternative, i.e. against the alternative $H_1 : m \neq m_0$, no best critical region is available. We may, however, consider, by a practical compromise of the above results, a symmetrical two-tailed test where the critical region consists of the pair of intervals $(-\infty, -u_\epsilon)$ and (u_ϵ, ∞) , u_ϵ being given by

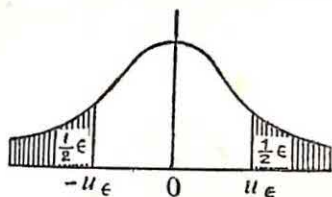


Fig. 31

$$P(U < -u_\epsilon) + P(U > u_\epsilon) = \epsilon$$

or

$$P(U > u_\epsilon) = \frac{1}{2}\epsilon \quad (16.4.4)$$

We may certainly expect that this will prove to be a good test.

Remark. In the last case for testing $H_0 : m = m_0$ against no alternative, the hypothesis will be accepted if

$$-u_\epsilon < u < u_\epsilon$$

where u_ϵ is given by (16.4.4), or

$$-u_\epsilon < \frac{\sqrt{n}(\bar{x} - m_0)}{\sigma} < u_\epsilon$$

or

$$\bar{x} - \frac{\sigma u_\epsilon}{\sqrt{n}} < m_0 < \bar{x} + \frac{\sigma u_\epsilon}{\sqrt{n}}$$

i.e. m_0 lies in the interval $\left(\bar{x} - \frac{\sigma u_\epsilon}{\sqrt{n}}, \bar{x} + \frac{\sigma u_\epsilon}{\sqrt{n}}\right)$ which, by (14.5.1) and (14.5.2), is nothing but the symmetrical confidence interval for m with confidence coefficient $1 - \epsilon$. This reconciliation is indeed interesting as well as logically satisfactory.

Example 1. In a ceramic industry the population of percentages of yield of first class material was known to have a mean 72.6 and standard deviation 2.4 (as obtained from past large samples). A new incentive bonus was declared, and a subsequent sample of size 15 from the population gave a mean of 74.3. Does this reasonably show that the bonus really helped raising the average yield percentage?

(It is known from experience that populations of yield percentages have normal or approximately normal distributions).

We assume that the population standard deviation remains unaltered, i.e. the new population is normal (m, σ) where $\sigma=2.4$. Since we are anticipating a rise in the average yield percentage or $m > 72.6$, we have to pose the null hypothesis

$$H_0 : m = 72.6 = m_0 \text{ (say)}$$

against the alternative

$$H_1 : m > m_0$$

Here the appropriate test will be a right-tailed standard normal test, and let us test at 5% level of significance.

For $\epsilon=.05$, u_ϵ is given by

$$P(U > u_\epsilon) = .05$$

whence, from Table I, $u_\epsilon=1.65$, so that the best critical region of the test is $u > 1.65$.

As $n=15$, $\bar{x}=74.3$, $\sigma=2.4$, $m_0=72.6$, the computed value of $u = \sqrt{n}(\bar{x} - m_0)/\sigma = 2.74$. Since this value of u falls within the critical region, we reject the null hypothesis that the mean yield percentage continues to be the same, so that our belief that the incentive bonus was really effective is confirmed.

If, however, we do not fix the significance level beforehand, we proceed as follows. The computed value of u is 2.74, and from Table I, $P(U > 2.74) = .0031$ which shows that the value of u falls within the critical region of even 1% level or, according to our terminology, the value of u is highly significant. Hence we can confidently (i.e. with small risk of a wrong decision) reject H_0 .

Remarks

1. We know that the sample mean is an estimate of the population mean. Now the mean of the sample from the new population is 74.3 which is greater than the mean of the old population 72.6; but we note that the difference between the two is not too marked to enable us to rush to the conclusion that the population mean has really increased, for this difference might also have arisen due to random fluctuations. Hence, in order to come to a definite conclusion, we must have to make a statistical test, as done above, which depends on, besides other factors, the size of the sample, the population standard deviation etc.

2. The above example is one of the many practical problems in which a hypothesis arises naturally as a null hypothesis.

Test for σ . Consider the hypothesis

$$H_0 : \sigma = \sigma_0$$

against an alternative

$$H_1 : \sigma = \sigma_1$$

Since m is unspecified both H_0 and H_1 are composite hypotheses. But these can be regarded as simple hypotheses if we consider the population of the statistic s^2 (or S^2), the density function of which, given by (13.4.2), contains the only parameter σ . When a sample of size n is drawn from the parent population and the value of s^2 is computed, we get a sample of unit size from the population of s^2 , for which the likelihood function is

$$L(s^2; \sigma) = f_{s^2}(s^2; \sigma) = \frac{1}{\Gamma(\frac{1}{2}v)} \left(\frac{v}{2\sigma^2} \right)^{v/2} (s^2)^{v/2-1} e^{-vs^2/2\sigma^2}$$

The form of the best critical region is then given by

$$\frac{L(s^2; \sigma_0)}{L(s^2; \sigma_1)} = \left(\frac{\sigma_1}{\sigma_0} \right)^v e^{-\frac{1}{2}nS^2(1/\sigma_0^2 - 1/\sigma_1^2)} < k \quad [nS^2 = vs^2]$$

Case I. $\sigma_1 > \sigma_0$. The above inequality may be written in the form

$$\chi^2 = \frac{nS^2}{\sigma_0^2} > \chi_{\epsilon}^2 \quad (16.4.5)$$

so that the required statistic may be chosen to be $\chi^2 = nS^2/\sigma_0^2$, the sampling distribution of which has, under H_0 , a χ^2 -distribution with $v = n - 1$ degrees of freedom, and the corresponding best critical region is the right tail $\chi^2 > \chi_{\epsilon}^2$ of the χ^2 -density curve, χ_{ϵ}^2 being given by

$$P(\chi^2 > \chi_{\epsilon}^2) = \epsilon \quad (16.4.6)$$

Since the statistic and its best critical region are independent of the particular value of σ_1 but depend only on the fact that $\sigma_1 > \sigma_0$, the test holds against the alternative $H_1 : \sigma > \sigma_0$.

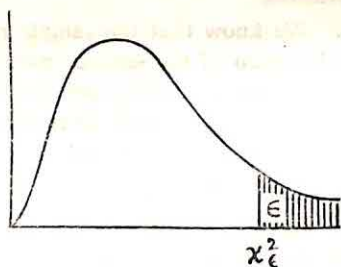


Fig. 32

Case II. $\sigma_1 < \sigma_0$. In this case we have

$$\chi^2 = \frac{nS^2}{\sigma_0^2} < \chi_{\epsilon}^2 \quad (16.4.7)$$

which shows that the best critical region is now the left tail $0 < \chi^2 < \chi^2_{\epsilon}$ where

$$P(0 < \chi^2 < \chi^2_{\epsilon}) = \epsilon$$

or

$$P(\chi^2 > \chi^2_{\epsilon}) = 1 - \epsilon \quad (16.4.8)$$

This gives the best test against the alternative $H_1: \sigma < \sigma_0$. In order to test $H_0: \sigma = \sigma_0$ against no alternative, it is customary to use a two-tailed χ^2 -test, in which the critical region consists of the intervals $(0, \chi^2_{\epsilon/2})$ and $(\chi^2_{\epsilon/2}, \infty)$ such that

$$P(0 < \chi^2 < \chi^2_{\epsilon/2}) = \frac{1}{2}\epsilon$$

or

$$P(\chi^2 > \chi^2_{\epsilon/2}) = 1 - \frac{1}{2}\epsilon$$

and

$$P(\chi^2 > \chi^2_{\epsilon/2}) = \frac{1}{2}\epsilon$$

(16.4.9)

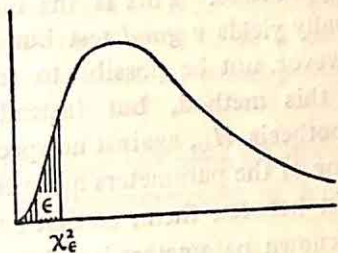


Fig. 33

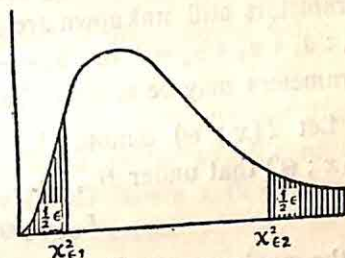


Fig. 34

Example 2. If in Ex. 1 the sample standard deviation was 1.9, test at 5% level if our previous assumption that the standard deviation of the new population continued to be 2.4 was justified.

We are required to test the null hypothesis $H_0: \sigma = 2.4 = \sigma_0$ (say), and as there is no specific alternative, the two-tailed χ^2 -test should be used.

For $\nu = 14$ degrees of freedom and $\epsilon = 0.05$, $\chi^2_{\epsilon/2}$ and $\chi^2_{\epsilon/2}$ are given by

$$P(\chi^2 > \chi^2_{\epsilon/2}) = .975, \quad P(\chi^2 > \chi^2_{\epsilon/2}) = .025$$

By Table II $\chi^2_{\epsilon/2} = 5.569$, $\chi^2_{\epsilon/2} = 26.342$. Hence the critical region consists of the intervals $(0, 5.569)$ and $(26.342, \infty)$.

Since $S = 1.9$, $\chi^2 = nS^2/\sigma_0^2 = 9.40$ which does not fall in the critical region, and hence we accept H_0 . It will be seen that the value of χ^2 is not significant even at 20% or 30% level, so that our assumption that the standard deviation of the new population was 2.4 seems quite reasonable.

16.5 LIKELIHOOD RATIO TESTING

When the hypotheses are composite or even when they are simple for which the Neyman-Pearson theorem does not lead to a convenient

test, we may take recourse to another method for constructing tests of hypotheses. This is the method of *likelihood ratio testing* which usually yields a *good* test but not necessarily the best test. It will, however, not be possible to treat any general composite hypothesis by this method, but instead we shall consider a composite null hypothesis H_0 , against no specific alternative, of the following forms : some of the parameters are specified or some given functional relations exist between them, so that under the hypothesis H_0 the number of unknown parameters is reduced. Let Θ' denote the set of parameters still unknown under H_0 . Suppose, for example, $\Theta = (\theta_1, \theta_2, \theta_3)$. (i) If $H_0 : \theta_1 = 1$, then under H_0 the parameter θ_1 becomes known, and the parameters still unknown are θ_2 and θ_3 , and hence $\Theta' = (\theta_2, \theta_3)$ (ii) If $H_0 : \theta_1 + \theta_2 + \theta_3 = 0$ or $\theta_1 = -\theta_2 - \theta_3$, then under H_0 the unknown parameters may be taken to be θ_2, θ_3 , i.e. $\Theta' = (\theta_2, \theta_3)$.

Let $L(x; \Theta)$ denote the likelihood function of the sample and $L_0(x; \Theta')$ that under H_0 , i.e.

$$L_0(x; \Theta') = L(x; \Theta | H_0) \quad (16.5.1)$$

If the maximum likelihood estimates $\hat{\Theta} = \hat{\Theta}(x)$ and $\hat{\Theta}' = \hat{\Theta}'(x)$ of Θ and Θ' respectively exist, then

$$\max L(x; \Theta) = L(x; \hat{\Theta}), \quad \max L_0(x; \Theta') = L_0(x; \hat{\Theta}')$$

or, speaking clearly, the maximum value of the likelihood function $L(x; \Theta)$ (for a fixed sample point) when the parametric point Θ is allowed to vary over the entire admissible part of the parametric space P_k is $L(x; \hat{\Theta})$, and the same when the parametric point is allowed to vary only over a set of points in P_k given by H_0 is $L_0(x; \hat{\Theta}')$. Obviously it follows that $L_0(x; \hat{\Theta}')$ cannot exceed $L(x; \hat{\Theta})$, i.e.

$$0 \leq L_0(x; \hat{\Theta}') \leq L(x; \hat{\Theta})$$

Setting

$$\lambda = \frac{L_0(x; \hat{\Theta}')}{L(x; \hat{\Theta})} \quad (16.5.2)$$

we have

$$0 \leq \lambda \leq 1 \quad (16.5.3)$$

The statistic $\lambda = \lambda(x)$ which is free from unknown parameters is called the *likelihood ratio* for H_0 , and (16.5.3) shows that the spectrum of the corresponding variate λ is the interval $(0, 1)$. Now the density function of λ under H_0 , in general, depends on θ' ; but if, in a particular case, it is independent of all unknown parameters, we can proceed to construct a test of H_0 as follows.

If H_0 is true, $L_0(x; \theta') = L(x; \theta)$, and since maximum likelihood estimates are known to be good estimates of the parameters, we have

$$L(x; \theta) \simeq L(x; \hat{\theta}), \quad L_0(x; \theta') \simeq L_0(x; \hat{\theta}')$$

and hence if H_0 is true, $L_0(x; \hat{\theta}') \simeq L(x; \hat{\theta})$ or $\lambda \simeq 1$. Thus if the observed value of the statistic λ lies close to 1, we may reasonably believe that H_0 is true. On the other hand, if the observed value of λ is close to 0, it follows that $L_0(x; \hat{\theta}') \ll L(x; \hat{\theta})$ or $L_0(x; \theta') \ll L(x; \theta)$ which shows that it is plausible to conclude that H_0 is false. Hence, for testing H_0 we can take λ as the statistic of the test, for which the critical region will be $(0, \lambda_e)$ where λ_e ($0 < \lambda_e < 1$) is a constant such that the probability of Type I error,

$$P(0 < \lambda < \lambda_e | H_0) = \int_0^{\lambda_e} f_\lambda(\lambda) d\lambda = \varepsilon \quad (16.5.4)$$

where $f_\lambda(\lambda)$ is the density function of λ under H_0 . Since $f_\lambda(\lambda)$ is assumed to be independent of unknown parameters, the equation (16.5.4) uniquely determines λ_e as a function of ε .

Remarks

1. The above method does not always yield a test, for it rests upon the assumption that the density function of λ under H_0 does not depend on unknown parameters which is not always the case.

2. The likelihood ratio test is based on *intuitive* concepts and not on any exact logical criterion. We have, in fact, disregarded specific alternatives and have not also taken direct account of the Type II error. It can, however, be proved by further investigations that the likelihood ratio test generally corresponds to relatively small Type II error and as such is a good test. When a specific alternative to the null hypothesis

is not stated, the best test usually does not exist, and in such cases the likelihood ratio test provides a good second choice.

3. We can make a transformation $z = z(\lambda)$, in case z is a more convenient statistic than λ . If the critical interval $0 < \lambda < \lambda_c$ transforms to the form $z \in R_c$ where R_c is a region of the z -axis, then R_c is the critical region for the statistic z determined by

$$P(Z \in R_c | H_0) = \int_{R_c} f_z(z) dz = \epsilon \quad (16.5.5)$$

where $f_z(z)$ is the density function of Z under H_0 , which is, of course, independent of unknown parameters.

16.6 NORMAL (m, σ) POPULATION

Test for m . $H_0: m = m_0$

Case I. σ known. H_0 is a simple hypothesis. From Sec. 14.2

$$L(x; m) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m)^2}$$

or

$$\log L = -\frac{1}{2\sigma^2} \sum (x_i - m)^2 + \text{terms independent of } m$$

The likelihood equation is $\frac{\partial \log L}{\partial m} = 0$ or $\sum (x_i - m) = 0$ which gives

$\hat{m} = \bar{x}$. So

$$L(x; \hat{m}) = (2\pi)^{-n/2} \sigma^{-n} e^{-nS^2/2\sigma^2}$$

The likelihood function under H_0 ,

$$L_0(x) = L(x; m_0) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m_0)^2}$$

Hence the likelihood ratio

$$\begin{aligned} \lambda &= \frac{L_0(x)}{L(x; \hat{m})} = e^{-\frac{1}{2\sigma^2} \sum [(x_i - m_0)^2 - nS^2]} \\ &= e^{-n(\bar{x} - m_0)^2/2\sigma^2} = e^{-u^2/2} \end{aligned}$$

where

$$u = \frac{\sqrt{n}(\bar{x} - m_0)}{\sigma} \quad (16.6.1)$$

Now instead of λ it will be more convenient to take u as our statistic whose sampling distribution under H_0 is known to be normal $(0, 1)$. The critical interval $0 < \lambda < \lambda_\epsilon$ clearly transforms to the form $|u| > u_\epsilon$, i.e. the two tails $(-\infty, -u_\epsilon)$ and (u_ϵ, ∞) of the standard normal density curve, where u_ϵ is given by

$$P(|U| > u_\epsilon) = \epsilon$$

or

$$P(U > u_\epsilon) = \frac{1}{2}\epsilon \quad (16.6.2)$$

Thus we arrive at the same test as given earlier by the Neyman-Pearson theorem together with a practical compromise (cf. Sec. 16.4).

Case II. σ unknown. Here the hypothesis H_0 is composite, and the likelihood ratio test will lead to a result typical of its own. The likelihood function is given by

$$L(x; m, \sigma) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m)^2}$$

We know $\hat{m} = \bar{x}$, $\hat{\sigma} = S$ so that

$$L(x; \hat{m}, \hat{\sigma}) = (2\pi)^{-n/2} S^{-n} e^{-n/2}$$

Now

$$L_0(x; \sigma) = L(x; m_0, \sigma) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m_0)^2}$$

or

$$\log L_0 = -n \log \sigma - \frac{1}{2\sigma^2} \sum (x_i - m_0)^2 + \text{const.}$$

$$\frac{\partial \log L_0}{\partial \sigma} = 0 \quad \text{gives} \quad -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (x_i - m_0)^2 = 0, \quad \text{or}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - m_0)^2 = S^2 + (\bar{x} - m_0)^2$$

So

$$L_0(x; \hat{\sigma}) = (2\pi)^{n/2} \left\{ S^2 + (\bar{x} - m_0)^2 \right\}^{-n/2} e^{-n/2}$$

Hence

$$\lambda = \frac{L_0(x; \hat{\sigma})}{L(x; \hat{m}, \hat{\sigma})} = \left\{ 1 + \frac{(\bar{x} - m_0)^2}{S^2} \right\}^{-n/2}$$

or

$$\lambda = \left(1 + \frac{t^2}{v} \right)^{-n/2} \quad [v = n - 1]$$

where

$$t = \frac{\sqrt{n}(\bar{x} - m_0)}{s} \quad (16.6.3)$$

We know that, under H_0 , the random variable t has a t -distribution with $v = n - 1$ degrees of freedom. Hence we choose t as the statistic of the test, and the critical region $0 < \lambda < \lambda_\epsilon$ corresponds to $|t| > t_\epsilon$ where

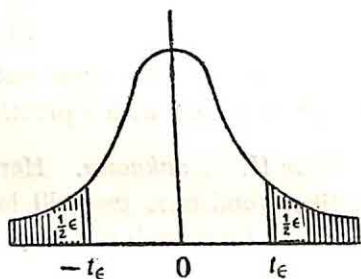


Fig. 35

$$P(|t| > t_\epsilon) = \epsilon$$

or

$$P(t > t_\epsilon) = \frac{1}{2}\epsilon \quad (16.6.4)$$

ONE-TAILED t -TESTS. To test $H_0 : m = m_0$ against the alternative $H_1 : m > m_0$, the intuitive modification of the above test will be a right-tailed t -test, i.e. the critical region will be $t > t_\epsilon$ where $P(t > t_\epsilon) = \epsilon$. Similarly, for testing $H_0 : m = m_0$ against $H_1 : m < m_0$ we use a left-tailed t -test, the critical region being $t < -t_\epsilon$ where $P(t < -t_\epsilon) = \epsilon$ or $P(t > t_\epsilon) = \epsilon$.

Example 1. In Ex. 1 Sec. 16.4 we can also do without the assumption regarding the standard deviation of the new population. In that case we have to make a right-tailed t -test, using the standard deviation of the sample.

The statistic $t = \sqrt{n}(\bar{x} - m_0)/s = 3.348$. It is seen from Table III that, corresponding to $v = 14$ degrees of freedom, the critical region for 1% significance level is $t > 2.624$. Hence the observed value of t is highly significant, and we confidently reject H_0 .

Test for $H_0: \sigma = \sigma_0$

We have

$$L(x; m, \sigma) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m)^2}$$

As before

$$L(x; \hat{m}, \hat{\sigma}) = (2\pi)^{-n/2} S^{-n} e^{-n/2}$$

$$L_0(x; m) = L(x; m, \sigma_0) = (2\pi)^{-n/2} \sigma_0^{-n} e^{-\frac{1}{2\sigma_0^2} \sum (x_i - m)^2}$$

Easily we find $\hat{m} = \bar{x}$, and

$$L_0(x; \hat{m}) = (2\pi)^{-n/2} \sigma_0^{-n} e^{-nS^2/2\sigma_0^2}$$

So

$$\lambda = \frac{L_0(x; \hat{m})}{L(x; \hat{m}, \hat{\sigma})} = e^{n/2} \left(\frac{S}{\sigma_0} \right)^n e^{-nS^2/2\sigma_0^2}$$

or

$$\lambda = \lambda(\chi^2) = \left(\frac{e}{n} \right)^{n/2} e^{-\chi^2/2} (\chi^2)^{n/2}$$

where

$$\chi^2 = \frac{nS^2}{\sigma_0^2} \quad (16.6.5)$$

is the required statistic, the sampling distribution of which under H_0 is χ^2 -distributed with $\nu = n - 1$ degrees of freedom.

For finding the critical region for χ^2 , we note that the inverse function of $\lambda = \lambda(\chi^2)$, $\chi^2 = \chi^2(\lambda)$ is a double-valued function of λ ; $\lambda = 0$ when $\chi^2 = 0$ and $\chi^2 \rightarrow \infty$, and the equation $\lambda(\chi^2) = \lambda_c$ has two solutions $\chi^2 = \chi_{\epsilon 1}^2, \chi_{\epsilon 2}^2$ such that

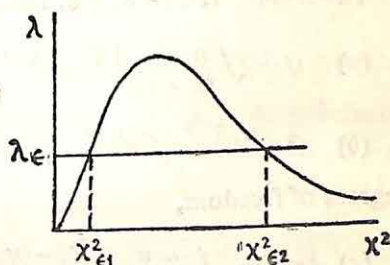


Fig. 36

$$\lambda(\chi_{\epsilon 1}^2) = \lambda(\chi_{\epsilon 2}^2)$$

(cf. Fig. 36). This equation determines $\chi_{\epsilon 2}^2$ as a function of $\chi_{\epsilon 1}^2$. Hence the critical region $(0, \lambda_c)$ for λ corresponds to the pair of

intervals $(0, \chi^2_{\epsilon_1})$ and $(\chi^2_{\epsilon_2}, \infty)$ for χ^2 where the only unknown $\chi^2_{\epsilon_1}$ is now found from the equation

$$P(\chi^2_{\epsilon_1} < \chi^2 < \chi^2_{\epsilon_2}) = 1 - \epsilon$$

The above method of determination of $\chi^2_{\epsilon_1}$, $\chi^2_{\epsilon_2}$ is, however, only theoretical and cannot be followed in practice. In practice, we usually make a two-tailed test such that the two tails of the χ^2 -density curve have equal areas, each being $\frac{1}{2}\epsilon$, i.e. $\chi^2_{\epsilon_1}$ and $\chi^2_{\epsilon_2}$ are given by (16.4.9). This was, in fact, the result obtained in Sec. 16.4.

16.7 COMPARISON OF NORMAL POPULATIONS

Let us be given two normal populations having parameters (m_1, σ_1) and (m_2, σ_2) , from which two independent samples

$$x = (x_1, x_2, \dots, x_i, \dots, x_{n_1})$$

of size n_1 and

$$x' = (x'_1, x'_2, \dots, x'_j, \dots, x'_{n_2})$$

of size n_2 are respectively drawn. On the basis of these samples, we shall test the equality of means and variances of the two populations. But before that, we prove that following results which will be necessary in the sequel. Let the characteristics of the first and second samples be marked with subscripts 1 and 2 respectively.

Theorem. If $\sigma_1 = \sigma_2 = \sigma$, the statistic

$$(a) \quad U = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sigma} \text{ is normal } (0, 1),$$

$$(b) \quad \chi^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \text{ is } \chi^2\text{-distributed with } v = n_1 + n_2 - 2$$

degrees of freedom,

$$(c) \quad t = \sqrt{v} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \text{ is } t\text{-distributed with } v = n_1 + n_2 - 2 \text{ degrees of freedom, and}$$

$$(d) \quad F = \frac{S_1^2}{S_2^2}, \text{ called the variance ratio, has an } F\text{-distribution with parameters } v_1 = n_1 - 1 \text{ and } v_2 = n_2 - 1.$$

Proof. We know that \bar{X}_1 and \bar{X}_2 are respectively normal $(m_1, \sigma/\sqrt{n_1})$ and $(m_2, \sigma/\sqrt{n_2})$. Since x and x' are independent, \bar{X}_1 and \bar{X}_2 are also so, and hence $\bar{X}_1 - \bar{X}_2$ is normal $(m_1 - m_2, \sigma\sqrt{(n_1 + n_2)/n_1 n_2})$ which immediately leads to (a).

Now $n_1 S_1^2 / \sigma^2$ and $n_2 S_2^2 / \sigma^2$ are independent χ^2 -variates with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom respectively, and hence their sum $(n_1 S_1^2 + n_2 S_2^2) / \sigma^2 = \chi^2$ is χ^2 -distributed with $v_1 + v_2 = n_1 + n_2 - 2 = v$ degrees of freedom which proves (b).

Since \bar{X}_1 and S_1^2 as well as \bar{X}_2 and S_2^2 are independent, it follows that U and χ^2 are independent, and by Theorem I Sec. 9.2 $\sqrt{v} U / \sqrt{\chi^2} = t$ has a t -distribution with v degrees of freedom. This is (c).

Again since $n_1 S_1^2 / \sigma^2$ and $n_2 S_2^2 / \sigma^2$ are independent χ^2 -variates with v_1 and v_2 degrees of freedom respectively, by Theorem I Sec. 9.3

$$\frac{v_2 n_1 S_1^2 / \sigma^2}{v_1 n_2 S_2^2 / \sigma^2} = \frac{S_1^2}{S_2^2} = F$$

has an F -distribution with parameters v_1, v_2 .

The likelihood functions of the two samples are respectively

$$L_1(x; m_1, \sigma_1) = (2\pi)^{-n_1/2} \sigma_1^{-n_1} e^{-\frac{1}{2\sigma_1^2} \Sigma(x_i - m_1)^2}$$

and

$$L_2(x'; m_2, \sigma_2) = (2\pi)^{-n_2/2} \sigma_2^{-n_2} e^{-\frac{1}{2\sigma_2^2} \Sigma(x'_j - m_2)^2}$$

so that their joint likelihood function (i.e. the joint density function of the independent random variables x and x') is given by

$$\begin{aligned} L(x, x'; m_1, m_2, \sigma_1, \sigma_2) &= L_1(x; m_1, \sigma_1) L_2(x'; m_2, \sigma_2) \\ &= (2\pi)^{-(n_1+n_2)/2} \sigma_1^{-n_1} \sigma_2^{-n_2} e^{-\frac{1}{2} \left[\frac{1}{\sigma_1^2} \Sigma(x_i - m_1)^2 + \frac{1}{\sigma_2^2} \Sigma(x'_j - m_2)^2 \right]} \quad (16.7.1) \end{aligned}$$

Test of equality of means. Assuming $\sigma_1 = \sigma_2 = \sigma$ (say), we shall construct a test for the hypothesis $H_0: m_1 = m_2$.

$$\begin{aligned} L(x, x'; m_1, m_2, \sigma) \\ = (2\pi)^{-(n_1+n_2)/2} \sigma^{-(n_1+n_2)} e^{-\frac{1}{2\sigma^2} [\Sigma(x_i - m_1)^2 + \Sigma(x'_j - m_2)^2]} \quad (16.7.2) \end{aligned}$$

Case I. σ known. We have

$$L(x, x'; m_1, m_2) = (2\pi)^{-(n_1+n_2)/2} \sigma^{-(n_1+n_2)} e^{-\frac{1}{2\sigma^2} [\Sigma(x_i - m_1)^2 + \Sigma(x_j' - m_2)^2]}$$

The likelihood equations are

$$\frac{\partial \log L}{\partial m_1} = 0, \quad \frac{\partial \log L}{\partial m_2} = 0$$

which respectively reduce to

$$\Sigma(x_i - m_1) = 0, \quad \Sigma(x_j' - m_2) = 0$$

giving

$$\hat{m}_1 = \bar{x}_1, \quad \hat{m}_2 = \bar{x}_2$$

So

$$L(x, x'; \hat{m}_1, \hat{m}_2) = (2\pi)^{-(n_1+n_2)/2} \sigma^{-(n_1+n_2)} e^{-(n_1 S_1^2 + n_2 S_2^2)/2\sigma^2}$$

The joint likelihood function under $H_0 : m_1 = m_2 = m$ (say),

$$L_0(x, x'; m) = L(x, x'; m, m)$$

$$= (2\pi)^{-(n_1+n_2)/2} \sigma^{-(n_1+n_2)} e^{-\frac{1}{2\sigma^2} [\Sigma(x_i - m)^2 + \Sigma(x_j' - m)^2]}$$

The equation $\frac{\partial \log L_0}{\partial m} = 0$ gives $\Sigma(x_i - m) + \Sigma(x_j' - m) = 0$ or

$$\hat{m} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

Now $\Sigma(x_i - \hat{m})^2 + \Sigma(x_j' - \hat{m})^2$

$$= \Sigma(x_i - \bar{x}_1)^2 + n_1(\bar{x}_1 - \hat{m})^2 + \Sigma(x_j' - \bar{x}_2)^2 + n_2(\bar{x}_2 - \hat{m})^2$$

$$= n_1 S_1^2 + n_2 S_2^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2$$

So

$$L_0(x, x'; \hat{m}) = (2\pi)^{-(n_1+n_2)/2} \sigma^{-\frac{1}{2\sigma^2} [n_1 S_1^2 + n_2 S_2^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2]}$$

The likelihood ratio

$$\lambda = \frac{L_0(x, x'; \hat{m})}{L(x, x'; \hat{m}_1, \hat{m}_2)} = e^{-n_2/2}$$

where

$$u = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \quad (16.7.3)$$

provides the suitable statistic for the test, the sampling distribution of which is normal $(0, 1)$ under H_0 . The critical region for λ is $0 < \lambda < \lambda_\epsilon$, and hence the same for u will be $|u| > u_\epsilon$, u_ϵ being given by

$$P(|U| > u_\epsilon) = \epsilon$$

or

$$P(U > u_\epsilon) = \frac{1}{2}\epsilon \quad (16.7.4)$$

For testing H_0 against the alternative $H_1 : m_1 > m_2$ or $m_1 < m_2$, we naturally consider one-tailed tests, a right-tailed test for the former and a left-tailed test for the latter.

Another method. We may suggest another simple method for deducing the above test. It is known that the population of the statistic $\bar{X}_1 - \bar{X}_2$ is normal $(m_1 - m_2, \sigma \sqrt{(n_1 + n_2)/n_1 n_2})$, so that $H_0 : m_1 = m_2$ is equivalent to the hypothesis that the mean of this population is zero. Since the computed value of $\bar{x}_1 - \bar{x}_2$ may be treated as a sample of size 1 from the corresponding population, the above test follows as a particular case of Sec. 16.6.

Case II. σ unknown. The joint likelihood function is given by (16.7.2). The likelihood equations are

$$\frac{\partial \log L}{\partial m_1} = 0, \quad \frac{\partial \log L}{\partial m_2} = 0$$

which respectively give

$$\hat{m}_1 = \bar{x}_1, \quad \hat{m}_2 = \bar{x}_2$$

and $\frac{\partial \log L}{\partial \sigma} = 0$ or

$$-\frac{n_1 + n_2}{\sigma} + \frac{1}{\sigma^3} \left\{ \sum (x_i - m_1)^2 + \sum (\bar{x}_j' - m_2)^2 \right\} = 0$$

giving

$$\hat{\sigma}^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

$$L(x, x'; \hat{m}_1, \hat{m}_2, \hat{\sigma}) \\ = (2\pi)^{-(n_1+n_2)/2} (n_1+n_2)^{(n_1+n_2)/2} (n_1 S_1^2 + n_2 S_2^2)^{-(n_1+n_2)/2} e^{-(n_1+n_2)/2}$$

Under $H_0 : m_1 = m_2 = m$ (say) the likelihood function becomes

$$L_0(x, x'; m, \sigma) = L(x, x'; m, m, \sigma) \\ = (2\pi)^{-(n_1+n_2)/2} \sigma^{-(n_1+n_2)} e^{-\frac{1}{2\sigma^2} [\Sigma(x_i - m)^2 + \Sigma(x_j' - m)^2]}$$

$$\frac{\partial \log L_0}{\partial m} = 0 \text{ gives } \Sigma(x_i - m) + \Sigma(x_j' - m) = 0 \quad \text{or}$$

$$\hat{m} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\text{and } \frac{\partial \log L_0}{\partial \sigma} = 0 \text{ gives}$$

$$-\frac{n_1 + n_2}{\sigma} + \frac{1}{\sigma^3} \left\{ \Sigma(x_i - m)^2 + \Sigma(x_j' - m)^2 \right\} = 0$$

or

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2} \left\{ \Sigma(x_i - \hat{m})^2 + \Sigma(x_j' - \hat{m})^2 \right\} \\ = \frac{1}{n_1 + n_2} \left\{ n_1 S_1^2 + n_2 S_2^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \right\}$$

$$L_0(x, x'; \hat{m}, \hat{\sigma}) = (2\pi)^{-(n_1+n_2)/2} (n_1 + n_2)^{(n_1+n_2)/2} \\ \times \left\{ n_1 S_1^2 + n_2 S_2^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \right\}^{-(n_1+n_2)/2} e^{-(n_1+n_2)/2}$$

Hence

$$\lambda = \frac{L_0(x, x'; \hat{m}, \hat{\sigma})}{L(x, x'; \hat{m}_1, \hat{m}_2, \hat{\sigma})}$$

or

$$\lambda = \left(1 + \frac{t^2}{v} \right)^{-(n_1+n_2)/2} \quad [v = n_1 + n_2 - 2]$$

where

$$t = \sqrt{v} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \quad (16.7.5)$$

Under H_0 the variate t has a t -distribution with ν degrees of freedom, and hence t can be chosen as the statistic of the test. The critical interval $0 < \lambda < \lambda_\epsilon$ corresponds to the critical region $|t| > t_\epsilon$ where

$$P(|t| > t_\epsilon) = \epsilon$$

or

$$P(t > t_\epsilon) = \frac{1}{2}\epsilon \quad (16.7.6)$$

A right-tailed test should be used for the alternative $H_1 : m_1 > m_2$, and a left-tailed test for the alternative $H_1 : m_1 < m_2$.

Example 1. There are two brands, A and B , of a type of string, of which the brand A sells at a slightly higher price than the brand B . 12 pieces of each brand of string were chosen at random, and their breaking strengths observed. The sample mean and variance for the A -strings were 18.8 lb and 4.08 lb², while those for the B -strings were 16.9 lb and 3.25 lb² respectively. Assuming that populations of breaking strengths of strings are normal and those of the A and B -strings have the same variance, test whether the A -strings are on the average better than the B -strings in respect of breaking strength.

Calling the populations of breaking strengths of the A and B -strings the first and second populations respectively, we set up the null hypothesis $H_0 : m_1 = m_2$ against the alternative $H_1 : m_1 > m_2$. For testing this a right-tailed t -test should be made, where the statistic t is given by (16.7.5).

Here $n_1 = n_2 = 12$, $\nu = 22$, $\bar{x}_1 = 18.8$, $S_1^2 = 4.08$, $\bar{x}_2 = 16.9$, $S_2^2 = 3.25$ which give $t = 2.328$. Now, for 22 degrees of freedom, the critical region for 5% level is $t > 1.717$, and that for 1% level is $t > 2.508$. These show that the value of t falls within the critical region of 5% level but outside that of 1% level, or, in other words, the value of t is significant. Hence we have some reasons for rejecting the null hypothesis and believing that the mean breaking strength of the A -strings is greater than that of the B -strings, although we are not very confident about it.

Test for equality of variances. $H_0 : \sigma_1 = \sigma_2$

The likelihood function is given by (16.7.1). Since $L = L_1 L_2$, we shall obviously get

$$\hat{m}_1 = \bar{x}_1, \hat{m}_2 = \bar{x}_2, \hat{\sigma}_1 = S_1, \hat{\sigma}_2 = S_2$$

$$L(x, x', \hat{m}_1, \hat{m}_2, \hat{\sigma}_1, \hat{\sigma}_2) = (2\pi)^{-(n_1+n_2)/2} S_1^{-n_1} S_2^{-n_2} e^{-(n_1+n_2)/2}$$

The joint likelihood function under $H_0 : \sigma_1 = \sigma_2 = \sigma$ (say),

$L_0(x, x'; m_1, m_2, \sigma)$ becomes identical with the likelihood function given by (16.7.2), and hence it follows from the preceding discussions that

$$L_0(x, x'; m_1, m_2, \sigma)$$

$$= (2\pi)^{-(n_1+n_2)/2} (n_1+n_2)^{(n_1+n_2)/2} (n_1 S_1^2 + n_2 S_2^2)^{-(n_1+n_2)/2} e^{-(n_1+n_2)/2}$$

Hence

$$\lambda = \lambda(F) = (n_1 + n_2)^{(n_1+n_2)/2} \left(1 - \frac{1}{n_1}\right)^{n_1/2} \left(1 - \frac{1}{n_2}\right)^{n_2/2} \times \frac{F^{n_1/2}}{(v_1 F + v_2)^{(n_1+n_2)/2}} \quad [v_1 = n_1 - 1, v_2 = n_2 - 1]$$

where

$$F = \frac{s_1^2}{s_2^2} \quad (16.7.7)$$

gives the required statistic, the sampling distribution of which under H_0 is known to be F -distributed with parameters v_1, v_2 .

Now $\lambda = 0$ when $F = 0$ as well as

$F \rightarrow \infty$, and the equation

$\lambda(F) = \lambda_\epsilon$ gives two solutions

$F = F_{\epsilon_1}, F_{\epsilon_2}$ such that $\lambda(F_{\epsilon_1})$

$= \lambda(F_{\epsilon_2})$ as shown in Fig. 37 so

that F_{ϵ_2} is obtained as a function

of F_{ϵ_1} . Therefore, the critical

interval $0 < \lambda < \lambda_\epsilon$ transforms to the two tails $0 < F < F_{\epsilon_1}$ and $F > F_{\epsilon_2}$ of the F -density curve, F_{ϵ_2} being finally determined by

$$P(F_{\epsilon_1} < F < F_{\epsilon_2}) = 1 - \epsilon$$

In practical problems, however, we simply consider equal area tails, i.e.

$$P(0 < F < F_{\epsilon_1}) = \frac{1}{2}\epsilon$$

or

$$\left. \begin{aligned} P(F > F_{\epsilon_1}) &= 1 - \frac{1}{2}\epsilon \\ \text{and} \\ P(F > F_{\epsilon_2}) &= \frac{1}{2}\epsilon \end{aligned} \right\} \quad (16.7.8)$$

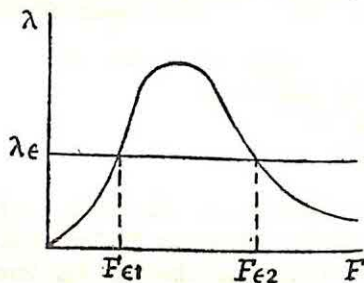


Fig. 37

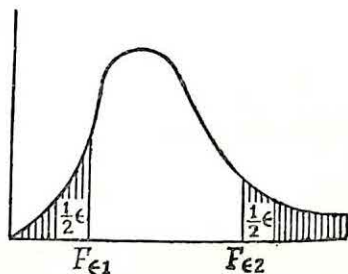


Fig. 38

Practical computation. The F -distribution, we note, depends on two parameters m, n , and as such preparation of detailed tables of the same becomes a difficult affair. Tables IV and V at the end of the book show only the points $F_{.15}$ and $F_{.01}$ respectively, where the notation F_ϵ is defined $P(F > F_\epsilon) = \epsilon$, corresponding to different values of m, n . These are respectively called the 5% and 1% points of the F -distribution. Thus for an F -test the right critical point can be directly obtained from the table (if, say, we want to test at 5% significance level, the right critical point is $F_{.025}$ which can be obtained from the values of $F_{.05}$ and $F_{.01}$ by the rule of proportional parts) but not the left critical point. The latter, however, can be calculated by making use of the fact that $1/F$ is also F -distributed having parameters n, m . For example, to find $F_{.95}$ we have

$$P(F > F_{.95}) = .95$$

or

$$P(F < F_{.95}) = .05$$

which is equivalent to

$$P\left(\frac{1}{F} > \frac{1}{F_{.95}}\right) = .05$$

so that $1/F_{.95}$ is the 5% point corresponding to the parameters n, m , from which we get $F_{.95}$.

We can, however, avoid calculating the left critical point altogether if we adopt the following procedure. We remember that for moderately large values of m, n the mode of the F -distribution occurs near about 1, and usually $F_{\epsilon_1} < 1$ and $F_{\epsilon_2} > 1$. Now if we place the larger s^2 in the numerator of the variance ratio (i.e. call the sample corresponding to the larger s^2 the first sample), the computed value of F will always be greater than 1 and can never fall in the left critical region, and hence it will suffice to calculate the right critical point F_{ϵ_2} only. This evidently means that instead of a two-tailed test at significance level ϵ , we are using a right-tailed test at level $\frac{1}{2}\epsilon$.

If we want to test $H_0 : \sigma_1 = \sigma_2$ against the alternative $H_1 : \sigma_1 > \sigma_2$, the proper test will, however, be a right-tailed test at the given significance level ϵ . For testing H_0 against the alternative $H_1 : \sigma_1 < \sigma_2$ a left-tailed test should be used, but this can again be

a reduced to a right-tailed test by simply interchanging the two populations. Thus in all cases we can conveniently stick to the right-tailed F -test.

Example 2. In the previous example test the assumption that the populations of breaking strengths of the A and B -strings have a common variance.

The null hypothesis is $H_0: \sigma_1 = \sigma_2$ against no alternative, which has to be tested by means of a two-tailed F -test.

The variances of the two samples are 4.08 and 3.25, and denoting the greater of these by S_1^2 we have $S_1^2 = 4.08$, $S_2^2 = 3.25$, so that $F = s_1^2/s_2^2 = 1.26$.

At 10% significance level the right critical point F_{ϵ_2} is given by $P(F > F_{\epsilon_2}) = .05$ corresponding to the parameters $\nu_1 = 11$, $\nu_2 = 11$. By Table IV $F_{\epsilon_2} = 2.83$, i.e. the right critical interval is $F > 2.83$ which shows that H_0 is accepted even at 10% level.

Remark. In Exs. 1 and 2 above, as also in many other problems, the population mean gives an index of the general or average quality of the products. The population standard deviation then provides a measure of variability or lack of uniformity of the quality. For good products the population naturally should have a large mean and a small standard deviation.

16.8 BIVARIATE NORMAL POPULATION

We shall here set up a test for the null hypothesis $H_0: \rho = 0$. For this, we will assume the following theorem whose deduction is somewhat complicated.

Theorem. If $\rho = 0$, the statistic $t = \sqrt{v} \frac{r}{\sqrt{1-r^2}}$ has a t -distribution with $v = n - 2$ degrees of freedom.

The likelihood function $L(x, y; m_x, m_y, \sigma_x, \sigma_y, \rho)$ is given by (15.5.4). Since $\hat{m}_x = \bar{x}$, $\hat{m}_y = \bar{y}$, $\hat{\sigma}_x = S_x$, $\hat{\sigma}_y = S_y$, $\hat{\rho} = r$, we have

$$L(x, y; \hat{m}_x, \hat{m}_y, \hat{\sigma}_x, \hat{\sigma}_y, \hat{\rho}) = (2\pi)^{-n} S_x^{-n} S_y^{-n} (1 - r^2)^{-n/2} e^{-n}$$

The likelihood function under H_0

$$\begin{aligned} L_0(x, y; m_x, m_y, \sigma_x, \sigma_y) \\ = (2\pi)^{-n} \sigma_x^{-n} \sigma_y^{-n} e^{-\frac{1}{2} \sum \left\{ \frac{(x_i - m_x)^2}{\sigma_x^2} + \frac{(y_i - m_y)^2}{\sigma_y^2} \right\}} \end{aligned}$$

Easily we obtain

$$\hat{m}_x = \bar{x}, \hat{m}_y = \bar{y}, \hat{\sigma}_x = S_x, \hat{\sigma}_y = S_y$$

So

$$L_0(x, y; \hat{m}_x, \hat{m}_y, \hat{\sigma}_x, \hat{\sigma}_y) = (2\pi)^{-n} S_x^{-n} S_y^{-n} e^{-n}$$

The likelihood ratio

$$\lambda = (1 - r^2)^{n/2} = \left(1 + \frac{t^2}{v}\right)^{-n/2} \quad [v = n - 2]$$

where

$$t = \sqrt{v} \frac{r}{\sqrt{1 - r^2}} \quad (16.8.1)$$

By the above theorem, the variate t under H_0 has a t -distribution with $v = n - 2$ degrees of freedom, and hence t is our required statistic. The critical region for t which corresponds to the critical region $0 < \lambda < \lambda_\epsilon$ for λ is $|t| > t_\epsilon$, where t_ϵ is given by

$$P(|t| > t_\epsilon) = \epsilon$$

or

$$P(t > t_\epsilon) = \frac{1}{2}\epsilon \quad (16.8.2)$$

In case we have an alternative $H_1: \rho > 0$ or $\rho < 0$, we have to use one-tailed t -tests, a right-tailed test for $H_1: \rho > 0$ and a left-tailed test for $H_1: \rho < 0$.

Example. A random sample of size 10 from a bivariate normal population is found to have a correlation coefficient 0.47. Test if the population can be regarded as uncorrelated.

$H_0: \rho = 0$, and the test is a two-tailed t -test, the statistic being given by (16.8.1).

Here $r = 0.47$, $v = 8$ so that $t = 1.506$. Now $P(|t| > 1.506) = .18$ which shows that the value of t is not at all significant. Hence we can reasonably regard the population as uncorrelated.

16.9 EXERCISES

1. For a normal (m, σ) population with known m , construct, by means of the Neyman-Pearson theorem, a test for the null hypothesis $H_0: \sigma = \sigma_0$ against the alternative $H_1: \sigma < \sigma_0$ or $\sigma > \sigma_0$.
2. Find, by the method of likelihood ratio testing, a test of $H_0: \sigma = \sigma_0$ for a normal (m, σ) population assuming that m is known.
3. In Ex. 11 Sec. 14.7 test at 5% significance level if the mean score of the population can be regarded as 15.5.

4. The percentage of carbon content of a certain variety of steel has a standard specification of .05. For 12 samples of this steel, the percentages of carbon content were found to have an average .0483 and standard deviation .00117. Do these data reasonably conform to the standard specification? (Assume that the population of percentages of carbon content is normal.)

5. A drug is given to 10 patients, and the increments in their blood pressure were recorded to be 3, 6, -2, 4, -3, 4, 6, 0, 0, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure? Test at 5% significance level, assuming the population to be normal.

6. In Ex. 13 Sec. 14.7 are there sufficient reasons to believe that the students of the given college are on the average less than 65 inches tall?

7. If the standard deviation of the sample cited in Ex. 11 Sec. 14.7 is 5.8, verify the information that the population standard deviation is 5.2.

8. 11 measured values of a physical quantity have a standard deviation 0.14. Is the suspicion that the standard deviation of the population of measured values (which is an inverse measure of precision of the measuring process) is greater than 0.1 true? Assume the population to be normal, and use 5% level of significance.

9. Independent samples of sizes 30 and 55 from two normal populations having a common variance 17.6 were found to have means 23.0 and 21.9 respectively. Test at 1% significance level whether the populations also have the same mean.

10. A sample of size 10 is drawn from each of two normal populations having the same variance which is unknown. If the mean and variance of the sample from the first population are 7 and 26 and those of the sample from the second population are 4 and 10, test at 5% significance level if the two populations have the same mean.

11. The IQ's of persons, chosen at random from each of two groups, tested by Terman Merrill (M-form) were as follows :

Group I	116	121	125	125	127	128	131	132	135	137
Group II	109	110	112	114	115	119	122	123	125	131

On the basis of these data can we reasonably believe that the persons of Group I have in general greater IQ than those of Group II? Assume that the populations of scores are normal and that they have a common variance. Confirm the latter hypothesis by means of an F -test.

12. For testing the effects of two types of fertilisers on the yield of wheat, 15 experimental plots of ground were available. Wheat was grown in these plots,

8 of which were treated with Fertiliser I and the remaining 7 with Fertiliser II, and the yields in kg. are given by the following table :

Fertiliser I	38.8	39.4	41.5	41.8	44.3	44.8	46.2	48.0
Fertiliser II	39.1	40.2	40.8	42.1	42.6	44.5	44.8	—

Do the two fertilisers really differ in their effects ? Assume that the two populations of yields of wheat, which can be taken to be normal, have the same variance. Also make a significance test of this hypothesis.

13. The lengths of life of 25 electric bulbs of one kind and 15 of another kind were found to have standard deviations 259 and 115 hours respectively. Test at 1% level of significance if the former kind of bulbs have less uniform quality than the latter, assuming that the populations in question are normal.

14. The correlation coefficient of a sample of size 5,000 from a bivariate normal population is $-.038$. Test the hypothesis $\rho=0$ against the alternative $\rho < 0$, where ρ denotes the correlation coefficient of the population.

15. In Ex. 2 Sec. 15.5 test if the bivariate population of theoretical and practical marks, which may be assumed to be normal, is at all correlated.

TESTING OF HYPOTHESES II

In this chapter we shall discuss some approximate tests, including what are known as tests for goodness of fit, by the method of likelihood ratio testing. Let us begin with the binomial population.

17.1 BINOMIAL (n, p) POPULATION

Let the parameter n be known and large and the null hypothesis to be tested be $H_0: p = p_0$. Let v denote an observed value of the binomial variate X , i.e. a sample of size 1 from the binomial population. The likelihood function is given by

$$L(v; p) = \binom{n}{v} p^v (1-p)^{n-v}$$

By Sec. 14.2, $\hat{p} = v/n$. Hence

$$L(v; \hat{p}) = \binom{n}{v} \left(\frac{v}{n}\right)^v \left(1 - \frac{v}{n}\right)^{n-v}$$

The likelihood function under H_0 ,

$$L_0(v) = L(v; p_0) = \binom{n}{v} p_0^v (1-p_0)^{n-v}$$

Hence

$$\lambda = \frac{L_0(v)}{L(v; \hat{p})} = \left(\frac{v}{np_0}\right)^{-v} \left(\frac{1-v/n}{1-p_0}\right)^{-n+v}$$

If H_0 is true, $v/n - p_0$ is a small quantity for large n , and we have

$$\lambda = \left(1 + \frac{v - np_0}{np_0}\right)^{-v} \left(1 - \frac{v - np_0}{nq_0}\right)^{-n+v} \quad [q_0 = 1 - p_0]$$

or

$$\begin{aligned} -\log \lambda &= v \log \left(1 + \frac{v - np_0}{np_0}\right) + (n - v) \log \left(1 - \frac{v - np_0}{nq_0}\right) \\ &= (v - np_0) \log \left(1 + \frac{v - np_0}{np_0}\right) + (n - v - nq_0) \log \left(1 - \frac{v - np_0}{nq_0}\right) \\ &\quad + np_0 \log \left(1 + \frac{v - np_0}{np_0}\right) + nq_0 \log \left(1 - \frac{v - np_0}{nq_0}\right) \end{aligned}$$

Noting that $n - v - np_0 = -(v - np_0)$,

$$\begin{aligned} -\log \lambda &\simeq \frac{(v - np_0)^2}{np_0} + \frac{(v - np_0)^2}{nq_0} + np_0 \left\{ \frac{v - np_0}{np_0} - \frac{(v - np_0)^2}{2n^2 p_0^2} \right\} \\ &\quad - nq_0 \left\{ \frac{v - np_0}{nq_0} + \frac{(v - np_0)^2}{2n^2 q_0^2} \right\} \\ &= \frac{(v - np_0)^2}{2np_0 q_0} \end{aligned}$$

or

$$-2 \log \lambda \simeq \frac{(v - np_0)^2}{np_0 q_0} = u^2$$

where

$$u = \frac{v - np_0}{\sqrt{np_0 q_0}} \quad (17.1.1)$$

When n is large, $U = \frac{X - np_0}{\sqrt{np_0 q_0}}$ is approximately normal $(0, 1)$ under H_0 , so that u can be taken to be the required statistic. We note that as λ ranges from 1 to 0, $-2 \log \lambda$ ranges from 0 to ∞ , and hence the critical region $0 < \lambda < \lambda_\epsilon$ for λ changes to $|u| > u_\epsilon$ for u where

$$P(|U| > u_\epsilon) = \epsilon$$

or

$$P(U > u_\epsilon) = \frac{1}{2}\epsilon \quad (17.1.2)$$

Remark. The approximation of $-2 \log \lambda$ under the hypothesis H_0 introduces a little crudeness in the logic of the process. This, in fact, increases the probability of Type II error, i.e. weakens the power of the test.

17.2 COMPARISON OF BINOMIAL POPULATIONS

Consider two binomial populations (n_1, p_1) and (n_2, p_2) , of which the parameters n_1 and n_2 are known and large, and let v_1 and v_2 be independent samples of unit size respectively from them. Our problem is to construct a test for the hypothesis $H_0: p_1 = p_2$. The joint likelihood function of two samples is given by

$$\begin{aligned} L(v_1, v_2; p_1, p_2) \\ = \binom{n_1}{v_1} \binom{n_2}{v_2} p_1^{v_1} p_2^{v_2} (1 - p_1)^{n_1 - v_1} (1 - p_2)^{n_2 - v_2} \end{aligned}$$

Easily we get $\hat{p}_1 = v_1/n_1$, $\hat{p}_2 = v_2/n_2$. So

$$L(v_1, v_2; \hat{p}_1, \hat{p}_2) \\ = \binom{n_1}{v_1} \binom{n_2}{v_2} \left(\frac{v_1}{n_1}\right)^{v_1} \left(\frac{v_2}{n_2}\right)^{v_2} \left(1 - \frac{v_1}{n_1}\right)^{n_1 - v_1} \left(1 - \frac{v_2}{n_2}\right)^{n_2 - v_2}$$

Under $H_0 : p_1 = p_2 = p$ (say) the likelihood function becomes

$$L_0(v_1, v_2; p) = \binom{n_1}{v_1} \binom{n_2}{v_2} p^{v_1 + v_2} (1 - p)^{n_1 + n_2 - v_1 - v_2}$$

which gives

$$\hat{p} = \frac{v_1 + v_2}{n_1 + n_2} \quad (17.2.1)$$

$$L_0(v_1, v_2; \hat{p}) = \binom{n_1}{v_1} \binom{n_2}{v_2} \hat{p}^{v_1 + v_2} (1 - \hat{p})^{n_1 + n_2 - v_1 - v_2}$$

Hence

$$\lambda = \left(\frac{v_1}{n_1 \hat{p}}\right)^{-v_1} \left(\frac{v_2}{n_2 \hat{p}}\right)^{-v_2} \left(\frac{1 - v_1/n_1}{1 - \hat{p}}\right)^{-n_1 + v_1} \left(\frac{1 - v_2/n_2}{1 - \hat{p}}\right)^{-n_2 + v_2}$$

Since n_1, n_2 are large, under H_0 , $\hat{p}_1 \simeq \hat{p}$ and $\hat{p}_2 \simeq p$, i.e. $v_1/n_1 - \hat{p}$ and $v_2/n_2 - \hat{p}$ are small quantities, and, as in the last section, we shall obtain

$$\begin{aligned} -2 \log \lambda &\simeq \frac{(v_1 - n_1 \hat{p})^2}{n_1 \hat{p} \hat{q}} + \frac{(v_2 - n_2 \hat{p})^2}{n_2 \hat{p} \hat{q}} \quad [\hat{q} = 1 - \hat{p}] \\ &= \frac{n_1 n_2}{\hat{p} \hat{q} (n_1 + n_2)} \left(\frac{v_1}{n_1} - \frac{v_2}{n_2}\right)^2 = u^2 \end{aligned}$$

where

$$u = \left(\frac{v_1}{n_1} - \frac{v_2}{n_2}\right) / \sqrt{\hat{p} \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (17.2.2)$$

Let v_1 and v_2 be the observed values of the binomial variates X_1 and X_2 respectively. If H_0 is true, X_1 and X_2 are approximately normal $(n_1 p, \sqrt{n_1 p q})$ and $(n_2 p, \sqrt{n_2 p q})$ respectively for large n_1, n_2 ($q = 1 - p$). Since v_1 and v_2 are obtained by independent observations, X_1, X_2 are independent, and hence $X_1/n_1 - X_2/n_2$ is approximately

normal $\left(0, \sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$. Now the parameter p which is unknown may be replaced as an approximation by its maximum likelihood estimate \hat{p} given by (17.2.1), so that under H_0 the sampling distribution of u is approximately standard normal. Hence, instead of λ , we can choose u as the statistic of the test, the critical region for which becomes $|u| > u_\epsilon$ where

$$P(|U| > u_\epsilon) = \epsilon$$

or

$$P(U > u_\epsilon) = \frac{1}{2}\epsilon \quad (17.2.3)$$

Example. In reply to the question : "Do you usually enjoy the evening alone" 96 out of 542 persons from one population and 117 out of 979 from another population said "yes". Do the two populations really have different psychological attitudes towards the given question ?

For the first population the number of 'yeses' in 542 trials is a binomial $(n_1=542, p_1)$ variate, an observed value of which is $v_1=96$, and the same for the second population in 979 trials is a binomial $(n_2=979, p_2)$ variate whose observed value is $v_2=117$. Clearly, the question requires to test the null hypothesis $H_0 : p_1 = p_2$.

By (17.2.2) $\hat{p} = .14004$, and hence $\hat{q} = .85996$, so that according to (17.2.3) $u = 3.10$.

Now $P(|U| > 3.10) = .002$ which shows that the value of the statistic u is highly significant, and the null hypothesis is rejected.

17.3. POISSON- μ POPULATION

$$H_0 : \mu = \mu_0$$

The likelihood function for a sample of size n ,

$$L(x; \mu) = e^{-n\bar{x}} \frac{\mu^{n\bar{x}}}{x_1! x_2! \dots x_n!}$$

We know $\hat{\mu} = \bar{x}$ so that

$$L(x; \hat{\mu}) = e^{-n\bar{x}} \frac{\bar{x}^{n\bar{x}}}{x_1! x_2! \dots x_n!}$$

Now

$$L_0(x) = L(x; \mu_0) = e^{-n\mu_0} \frac{\mu_0^{n\bar{x}}}{x_1! x_2! \dots x_n!}$$

So

$$\lambda = e^{n(\bar{x} - \mu_0)} \left(\frac{\bar{x}}{\mu_0} \right)^{-n\bar{x}} = e^{n(\bar{x} - \mu_0)} \left(1 + \frac{\bar{x} - \mu_0}{\mu_0} \right)^{-n\bar{x}}$$

or

$$\begin{aligned} -\log \lambda &= -n(\bar{x} - \mu_0) + n\bar{x} \log \left(1 + \frac{\bar{x} - \mu_0}{\mu_0} \right) \\ &= -n(\bar{x} - \mu_0) + n(\bar{x} - \mu_0) \log \left(1 + \frac{\bar{x} - \mu_0}{\mu_0} \right) \\ &\quad + n\mu_0 \log \left(1 + \frac{\bar{x} - \mu_0}{\mu_0} \right) \end{aligned}$$

For large n , $\bar{x} - \mu_0$ is a small quantity under H_0 , and hence

$$-2 \log \lambda \simeq \frac{n(\bar{x} - \mu_0)^2}{\mu_0} = u^2$$

where

$$u = \sqrt{\frac{n}{\mu_0}} (\bar{x} - \mu_0) \quad (17.3.1)$$

is the required statistic, the sampling distribution of which is approximately normal $(0, 1)$ under H_0 . As before, here also we get a two-tailed standard normal test.

Remark. We note here that the sampling distribution of $-2 \log \lambda$ under H_0 is approximately chi-square with 1 degree of freedom for large n . This is, however, a particular case of a general theorem which states that if the population distribution function satisfies certain regularity conditions, the sampling distribution of $-2 \log \lambda$ under the null hypothesis is approximately chi-square for large samples.

17.4 MULTINOMIAL DISTRIBUTION

The probability distribution corresponding to the multinomial law is called the multinomial distribution which is a generalisation of the binomial distribution. Following the terminology of Sec. 4.6, let X_k denote the frequency of the event point U_k in the sequence E_n of n independent repetitions of E ($k = 1, 2, \dots, m$). Then X_1, X_2, \dots, X_m are random variables defined on the space S^n which satisfy the relation

$$\sum X_k = n \quad (17.4.1)$$

The joint distribution of the variables X_1, X_2, \dots, X_m , i.e. the distribution of the variable (X_1, X_2, \dots, X_m) subject to the constraint (17.4.1) is called the *multinomial distribution*. The spectrum of the multinomial distribution then consists of the points (i_1, i_2, \dots, i_m) where $i_1, i_2, \dots, i_m = 0, 1, 2, \dots, n$ such that $\sum i_k = n$, and the probability masses are given by

$$\begin{aligned} f_{i_1 i_2 \dots i_m} &= P(X_1 = i_1, X_2 = i_2, \dots, X_m = i_m) \\ &= \frac{n!}{i_1! i_2! \dots i_m!} p_1^{i_1} p_2^{i_2} \dots p_m^{i_m} \quad (17.4.2) \end{aligned}$$

which follows immediately from (4.6.3).

1. The multinomial distribution is $(m-1)$ -dimensional, the spectrum being confined to the $(m-1)$ -dimensional hyperplane $\sum x_k = n$ in the m -dimensional (x_1, x_2, \dots, x_m) -space. For $m=2$, we get the binomial distribution which is one-dimensional.

2. The multinomial distribution has m parameters p_1, p_2, \dots, p_m (apart from n) which are subject to $\sum p_k = 1$.

3. X_k is binomial (n, p_k) ($k=1, 2, \dots, m$).

Example. For n independent throws with a die, the joint distribution of the frequencies of the six different faces is a multinomial distribution with parameters $p_k = 1/6$ ($k=1, 2, \dots, 6$) and n .

We now state (without proof) an important theorem which is equivalent to a generalisation of the DeMoivre-Laplace Limit Theorem.

Theorem. If $n \rightarrow \infty$ (p_1, p_2, \dots, p_m being kept fixed), then the distribution function of

$$\sum \frac{(X_k - np_k)^2}{np_k} \quad (17.4.3)$$

tends to the χ^2 -distribution function with $\nu = m - 1$ degrees of freedom.

17.5 MULTINOMIAL POPULATION

Returning to statistical terminology, we now consider the population of the multinomial variate (X_1, X_2, \dots, X_m) . When the experiment E is repeated n times under uniform conditions (i.e. the compound experiment E_n is performed once), let the counted frequency of U_k be v_k ($k=1, 2, \dots, m$) so that (v_1, v_2, \dots, v_m) is an observed value of the random

variable (X_1, X_2, \dots, X_m) or, in other words, a sample of unit size from the corresponding multinomial population.

Estimation of parameters. Let us estimate the parameters p_k 's from this sample of unit size by the method of maximum likelihood, assuming n to be known. We have

$$L(v_1, v_2, \dots, v_m; p_1, p_2, \dots, p_m) = \frac{n!}{v_1! v_2! \dots v_m!} p_1^{v_1} p_2^{v_2} \dots p_m^{v_m}$$

or

$$\log L = \sum v_k \log p_k + \text{terms independent of } p_k\text{'s}$$

Here we shall have to maximise $\log L$ subject to the relation $\sum p_k = 1$, and hence

$$\sum \frac{v_k}{p_k} dp_k = 0, \quad \sum dp_k = 0$$

which give

$$\sum \left(\frac{v_k}{p_k} + \lambda \right) dp_k = 0$$

where λ is a Lagrange's multiplier. It follows that

$$\frac{v_k}{p_k} + \lambda = 0$$

for all k . Hence

$$\frac{v_1}{p_1} = \frac{v_2}{p_2} = \dots = \frac{v_m}{p_m} = \frac{\sum v_k}{\sum p_k} = n$$

or

$$\hat{p}_k = v_k/n \quad (k=1, 2, \dots, m)$$

which is an expected result.

Test of hypothesis. $H_0: p_k = p_{0k} \quad (k=1, 2, \dots, m)$ where p_{0k} are given positive numbers such that $\sum p_{0k} = 1$, and n (known) is large. Now

$$\begin{aligned} L(v_1, v_2, \dots, v_m; \hat{p}_1, \hat{p}_2, \dots, \hat{p}_m) \\ = \frac{n!}{v_1! v_2! \dots v_m!} \left(\frac{v_1}{n} \right)^{v_1} \left(\frac{v_2}{n} \right)^{v_2} \dots \left(\frac{v_m}{n} \right)^{v_m} \end{aligned}$$

and

$$L_0(v_1, v_2, \dots, v_m) = \frac{n!}{v_1! v_2! \dots v_m!} p_{01}^{v_1} p_{02}^{v_2} \dots p_{0m}^{v_m}$$

So

$$\lambda^{-1} = \left(\frac{v_1}{np_{01}}\right)^{v_1} \left(\frac{v_2}{np_{02}}\right)^{v_2} \dots \left(\frac{v_m}{np_{0m}}\right)^{v_m}$$

or

$$\begin{aligned} -\log \lambda &= \sum v_k \log \left(\frac{v_k}{np_{0k}}\right) \\ &= \sum \left[(v_k - np_{0k}) \log \left(1 + \frac{v_k - np_{0k}}{np_{0k}}\right) \right. \\ &\quad \left. + np_{0k} \log \left(1 + \frac{v_k - np_{0k}}{np_{0k}}\right) \right] \end{aligned}$$

Under H_0 each $v_k/n - p_{0k}$ is a small quantity for large n , and hence

$$-\log \lambda \simeq \sum \left[\frac{(v_k - np_{0k})^2}{np_{0k}} + np_{0k} \left\{ \frac{v_k - np_{0k}}{np_{0k}} - \frac{(v_k - np_{0k})^2}{2n^2 p_{0k}^2} \right\} \right]$$

or since $\sum (v_k - np_{0k}) = 0$,

$$-2\log \lambda \simeq \sum \frac{(v_k - np_{0k})^2}{np_{0k}} = \chi^2$$

where

$$\chi^2 = \sum \frac{(v_k - np_{0k})^2}{np_{0k}} \quad (17.5.1)$$

By the theorem of the previous section the sampling distribution of χ^2 under H_0 is approximately chi-square with $v = m - 1$ degrees of freedom for large n . Thus the statistic of the test is χ^2 ; the critical region $0 < \lambda < \lambda_\epsilon$ for λ corresponds to right tail $\chi^2 > \chi_\epsilon^2$ of the χ^2 -density curve where

$$P(\chi^2 > \chi_\epsilon^2) = \epsilon \quad (17.5.2)$$

Remark. The statistic χ^2 defined by (17.5.1) has an important significance. Consider the given event space S which contains the m points U_1, U_2, \dots, U_m . Now we can conceive of a statistical image of the probability distribution in S on the results of n repetitions of E by assigning a probability mass $1/n$ to each observed event point, so that, since v_k is the frequency of U_k , the latter gets of a share of mass v_k/n . On the other hand, in the theoretical distribution U_k carries a mass p_{0k} under the hypothesis H_0 , and hence any statistic of the form $\sum c_k^2 \left(\frac{v_k}{n} - p_{0k}\right)^2$ where c_k 's are any suitable constants, of which χ^2

defined by (17.5.1) is one, gives a *measure of deviation* of the empirical distribution from the assumed distribution. If H_0 is true, this measure should be small for large n , i.e. a large observed value of this measure will be an evidence against H_0 . This is in accord with the choice of the right-tail of the χ^2 -density curve as the critical region for the above test.

Example. In a cross-breeding experiment with plants of a certain species, 240 offspring were classified into 4 classes with respect to the structure of their leaves as follows :

Class	I	II	III	IV	Total
Frequency	127	40	52	21	240

According to Mendel's theory of heredity, the probabilities of the four classes should be in the ratio 9 : 3 : 3 : 1. Are these data consistent with the theory ?

Here we are concerned with a multinomial population with 4 parameters p_1, p_2, p_3, p_4 and have to test the hypothesis $H_0 : p_k = p_{0k} (k=1, 2, 3, 4)$ where

$$p_{01} = 9/16, p_{02} = 3/16, p_{03} = 3/16, p_{04} = 1/16$$

Class	ν	np_0	$(\nu - np_0)^2 / np_0$
I	127	135	0.474
II	40	45	0.556
III	52	45	1.089
IV	21	15	2.400
Total	240	240	4.519

Hence $\chi^2 = 4.519$, and for 3 degrees of freedom $P(\chi^2 > 4.519) = .21$ so that the value of χ^2 is not significant at all, and we may accept the hypothesis H_0 . In other words, the above data may be regarded to be consistent with the theory.

17.6 χ^2 -TEST OF GOODNESS OF FIT

We are now in a position to construct a test for a hypothesis of the type $H_0 : F(x) = F_0(x)$, where $F_0(x)$ is a given distribution function, on the basis of a large sample from the population. In other words, here we want to test if an observed sample fits an assumed population distribution, and hence such a test is customarily called a test of *goodness of fit*. Now the following two cases arise.

Case I. $F_0(x)$ is completely specified, i.e. contains no unknown parameters.

We propose to model this test after the χ^2 -test for a multinomial population discussed in the preceding section. For this, all that we have to do is to approximately reduce the given null hypothesis H_0 to a simple hypothesis regarding the parameters of some hypothetical multinomial population, which is done as follows :

1. Let X be the parent variable connected with the experiment E . We divide the spectrum of X into a finite number, say m , of suitable groups or classes C_1, C_2, \dots, C_m , having no common points (for a continuous distribution these will be in the form of intervals and, for the discrete case, in the form of groups of points of the spectrum), call the event $X \in C_k, U_k$, and set

$$p_k = P(U_k) = P(X \in C_k) \quad (k=1, 2, \dots, m) \quad (17.6.1)$$

Now noting that the events U_1, U_2, \dots, U_m are mutually exclusive and exhaustive, i.e. $U_1 + U_2 + \dots + U_m = S$, we may roughly regard U_1, U_2, \dots, U_m as the event points of S .

2. If X_k denotes the frequency of U_k in n independent repetitions of E ($k=1, 2, \dots, m$), then we know that (X_1, X_2, \dots, X_m) is a multinomial variate. An actual sample of size n (given by n repetitions of E) being drawn from the population of X , we count the number v_k of the sample values belonging to the class C_k ($k=1, 2, \dots, m$) and thereby obtain a sample of unit size, (v_1, v_2, \dots, v_m) from the multinomial population having parameters p_1, p_2, \dots, p_m .

3. Setting

$$p_{0k} = P(X \in C_k | H_0) \quad (k=1, 2, \dots, m) \quad (17.6.2)$$

which can be exactly calculated from $F_0(x)$, the hypothesis H_0 may be regarded to be approximately equivalent to the hypothesis H_0 : $p_k = p_{0k}$ ($k=1, 2, \dots, m$), and the latter can now be subjected to a χ^2 -test described in Sec. 17.5.

Case II. $F_0(x)$ has a known functional form but contains a number of unknown parameters $\theta_1, \theta_2, \dots, \theta_k$. This is the case which we usually encounter in practice, e.g. we wish to test if a population is normal or binomial etc.

In this case we first replace the parameters $\theta_1, \theta_2, \dots, \theta_k$ by their maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ respectively in $F_0(x)$ so that $F_0(x)$ becomes completely known. It can be proved (the proof being omitted) that if $F_0(x)$ satisfies certain general conditions, we may now employ the same procedure as in *Case I* with the only modification that the number of degrees of freedom of χ^2 is reduced by k corresponding to k parameters estimated from the sample, i.e. $\nu = m - k - 1$.

The approximation involved in the χ^2 -test holds only if n is large. For practical problems, it is usually found that the approximation is fairly good if $n > 50$. The approximation also depends upon the fact that each X_k , which is binomial (n, p_{0k}) under H_0 , is approximately normally distributed. This normal approximation, we remember, will not be valid if p_{0k} is very small, and it has been suggested for common problems that for the proper validity of the above test each *expected frequency* np_{0k} should be at least 5. As such if some of the expected frequencies are small, the χ^2 -test will be liable to error. We can, however, avoid this difficulty by combining together two or more adjacent small frequency classes to form a class for which the expected frequency is sufficiently large, i.e. exceeds 5.

Examples

1. For the data of Ex. 1 Sec. 12.3, can the population of numbers of α -ray counts be regarded as having a Poisson distribution?

The parameter μ of the Poisson distribution is unknown and is replaced by its estimate $\hat{\mu} = \bar{x}$ for calculating the expected frequencies. The last two entries of the given table are combined together, i.e. the Poisson spectrum of all non-negative integers is divided into 15 groups given by

$$i = 0, 1, 2, \dots, 13 \text{ and } \geq 14$$

so that the expected frequency np_{0k} for each group exceeds 5.

Thus if

$$f_i = e^{-\bar{x}} \frac{\bar{x}^i}{i!}$$

$$p_{0k} = f_k \quad (k = 0, 1, 2, \dots, 13)$$

$$= 1 - \sum_{i=0}^{13} f_i \quad (k = 14)$$

By the result of Ex. 1 Sec. 12.5, $\bar{x}=5.88$, and the computation is carried out as follows.

i	p_0	np_0	v	$(v - np_0)^2 / np_0$
0	.002795	9.7	8	0.298
1	.016433	56.8	59	0.085
2	.018314	166.9	177	0.611
3	.094696	327.2	311	0.802
4	.139203	480.9	492	0.256
5	.163702	565.6	528	2.500
6	.160428	554.3	601	3.934
7	.134760	465.6	467	0.004
8	.099045	342.2	331	0.367
9	.064712	223.6	220	0.058
10	.038050	131.5	121	0.838
11	.020340	70.3	85	3.074
12	.009966	34.4	24	3.144
13	.004508	15.6	22	2.626
≥ 14	.003048	10.5	9	0.214
Total	1.000000	3455.1	3455	18.811

Therefore $\chi^2=18.811$, and the number of degrees of freedom of χ^2 is 13 (remembering that the parameter μ has been estimated from the sample), for which $P(\chi^2 > 18.811) = .14$. This shows that the fit is quite satisfactory.

2. Can the rainfall data given in Ex. 2 Sec. 12.3 be regarded as a sample from a normal population?

The parameters m and σ of the assumed normal population have to be replaced by their estimates $\hat{m}=\bar{x}$ and $\hat{\sigma}=S$ respectively, so that the population distribution becomes completely known. The general procedure will be as follows.

The spectrum $(-\infty, \infty)$ of the normal distribution is divided into m suitable intervals $a_{k-1} < X \leq a_k$ ($k=1, 2, \dots, m$) which are our required classes. Obviously, the first class limit $a_0 = -\infty$ and the last $a_m = \infty$. Also it is necessary that the expected frequency of each of these classes is at least 5. Hence

In this case we first replace the parameters $\theta_1, \theta_2, \dots, \theta_k$ by their maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ respectively in $F_0(x)$ so that $F_0(x)$ becomes completely known. It can be proved (the proof being omitted) that if $F_0(x)$ satisfies certain general conditions, we may now employ the same procedure as in Case I with the only modification that the number of degrees of freedom of χ^2 is reduced by k corresponding to k parameters estimated from the sample, i.e. $\nu = m - k - 1$.

The approximation involved in the χ^2 -test holds only if n is large. For practical problems, it is usually found that the approximation is fairly good if $n > 50$. The approximation also depends upon the fact that each X_k , which is binomial (n, p_{0k}) under H_0 , is approximately normally distributed. This normal approximation, we remember, will not be valid if p_{0k} is very small, and it has been suggested for common problems that for the proper validity of the above test each *expected frequency* np_{0k} should be at least 5. As such if some of the expected frequencies are small, the χ^2 -test will be liable to error. We can, however, avoid this difficulty by combining together two or more adjacent small frequency classes to form a class for which the expected frequency is sufficiently large, i.e. exceeds 5.

Examples

1. For the data of Ex. 1 Sec. 12.3, can the population of numbers of α -ray counts be regarded as having a Poisson distribution?

The parameter μ of the Poisson distribution is unknown and is replaced by its estimate $\hat{\mu} = \bar{x}$ for calculating the expected frequencies. The last two entries of the given table are combined together, i.e. the Poisson spectrum of all non-negative integers is divided into 15 groups given by

$$i = 0, 1, 2, \dots, 13 \text{ and } \geq 14$$

so that the expected frequency np_{0k} for each group exceeds 5.

Thus if

$$f_i = e^{-\bar{x}} \frac{\bar{x}^i}{i!}$$

$$p_{0k} = f_k \quad (k = 0, 1, 2, \dots, 13)$$

$$= 1 - \sum_{i=0}^{13} f_i \quad (k = 14)$$

By the result of Ex. 1 Sec. 12.5, $\bar{x}=5.88$, and the computation is carried out as follows.

i	p_0	np_0	v	$(v - np_0)^2 / np_0$
0	.002795	9.7	8	0.298
1	.016433	56.8	59	0.085
2	.018314	166.9	177	0.611
3	.094696	327.2	311	0.802
4	.139203	480.9	492	0.256
5	.163702	565.6	528	2.500
6	.160428	554.3	601	3.934
7	.134760	465.6	467	0.004
8	.099045	342.2	331	0.367
9	.064712	223.6	220	0.058
10	.038050	131.5	121	0.838
11	.020340	70.3	85	3.074
12	.009966	34.4	24	3.144
13	.004508	15.6	22	2.626
≥ 14	.003048	10.5	9	0.214
Total	1.000000	3455.1	3455	18.811

Therefore $\chi^2=18.811$, and the number of degrees of freedom of χ^2 is 13 (remembering that the parameter μ has been estimated from the sample), for which $P(\chi^2 > 18.811) = .14$. This shows that the fit is quite satisfactory.

2. Can the rainfall data given in Ex. 2 Sec. 12.3 be regarded as a sample from a normal population?

The parameters m and σ of the assumed normal population have to be replaced by their estimates $\hat{m}=\bar{x}$ and $\hat{\sigma}=S$ respectively, so that the population distribution becomes completely known. The general procedure will be as follows.

The spectrum $(-\infty, \infty)$ of the normal distribution is divided into m suitable intervals $a_{k-1} < X \leq a_k$ ($k=1, 2, \dots, m$) which are our required classes. Obviously, the first class limit $a_0 = -\infty$ and the last $a_m = \infty$. Also it is necessary that the expected frequency of each of these classes is at least 5. Hence

$$p_{0k} = P(a_{k-1} < X \leq a_k)$$

where X denotes the parent random variable. Now the variable $Z = (X - \bar{x})/S$ is standard normal, and the corresponding standardised class limits z_k are given by

$$z_k = \frac{a_k - \bar{x}}{S} \quad (k=0, 1, 2, \dots, m)$$

whence

$$p_{0k} = P(z_{k-1} < Z \leq z_k) = \Phi(z_k) - \Phi(z_{k-1})$$

where $\Phi(z)$ is the standard normal distribution function.

In the present example, we combine the first 4 and the last 5 classes given in the original table, i.e. our classes are taken to be

$$(-\infty, 15), (15, 17), \dots, (25, 27), (27, \infty)$$

so that all the expected frequencies are greater than 5. From Ex. 2 Sec. 12.5, $\bar{x} = 21.157$, $S = 4.880$. The computation is shown below.

a	z	$\Phi(z)$	p_0	np_0	v	$(v - np_0)^2 / np_0$
$-\infty$	$-\infty$	0.0000				
15	-1.26	0.1038	0.1038	8.61	11	0.663
17	-0.85	0.1977	0.0939	7.79	5	0.999
19	-0.44	0.3300	0.1323	10.98	9	0.357
21	-0.03	0.4880	0.1580	13.11	9	1.288
23	0.38	0.6480	0.1600	13.28	21	4.488
25	0.79	0.7852	0.1372	11.39	13	0.228
27	1.20	0.8849	0.0997	8.28	6	0.628
∞	∞	1.0000	0.1151	9.56	9	0.033
Total	—	—	1.0000	83.00	83	8.684

Hence $\chi^2 = 8.684$. Since two parameters have been replaced by their estimates and there are 8 classes, χ^2 has 5 degrees of freedom, corresponding to which $P(\chi^2 > 8.684) = .13$. This shows that the hypothesis of normal population may be reasonably accepted.

17.7 EXERCISES

- 216 sixes were obtained in 1,000 throws with a die. Is the die honest?
- Let A be an event connected with a random experiment E . If in 192 repetitions of E under identical conditions A occurs 61 times, can we reasonably conclude that the probability of A is $\frac{1}{4}$? Use 5% level of significance.
- In Ex. 14 Sec. 14.7 is the belief that more than half of the electorate will vote in favour of the candidate reasonable?

4. Of 400 mangoes selected at random from a large stock, 53 were found to be bad. Test at 1% significance level the hypothesis that on the average 10% of the mangoes were bad.

5. In 80 tosses with one coin heads were obtained 27 times, and in 96 tosses with another coin heads were obtained 31 times. Show that both the coins may be regarded as biased. Are the coins equally biased?

6. In random samples of 374 and 210 persons from the adult populations of two large cities 72.4% and 88.1% were respectively found to be literate. Do the two populations really differ in their percentages of literacy?

7. In Ex. 4 Sec. 12.6 test if the mean number of daily telephone calls may be taken to be 8, assuming that the corresponding population has a Poisson distribution.

8. The number of daily accidents on a particular road was observed for 156 days, and the mean was found to be 0.165. Is the observed data consistent with the hypothesis that the average frequency of accidents is once in every 5 days? (Assume that the population in question is Poissonian.)

9. A die was thrown 1,000 times, and the frequencies of the different faces were observed to be the following :

Face	1	2	3	4	5	6	Total
Frequency	105	143	181	157	198	216	1,000

Test if the die is honest.

10. Of 160 offspring of a certain cross between guinea pigs, 102 were found to be red, 24 black and 34 white. According to a genetic model the probabilities of red, black and white are $9/16$, $3/16$ and $1/4$ respectively. Test at 5% significance level if the data are consistent with the model.

11. A random experiment has three outcomes— A , B , C which are exhaustive and mutually exclusive. The experiment was repeated 500 times, in which A , B , C respectively occurred 54, 258, 188 times. Test the hypothesis that the probabilities of A , B , C are in the ratio 1 : 5 : 4.

12. For the data given in Ex. 3 Sec. 12.6 test the hypotheses that the population of numbers of sixes is binomial ($5, p$) where (a) $p = 1/6$ (which serves as a test for the correctness of the die) and (b) p is unspecified.

13. Test if the population corresponding to the data of Ex. 4 Sec. 12.6 is Poissonian.

14. Test if the population corresponding to the data of Ex. 5 Sec. 12.6 is normal.

THEORY OF ERRORS

18.1 INTRODUCTION

The theory of errors was developed by Gauss, Laplace and others in the early 19th century long before modern statistics came into being. It was subsequently found that in the theory of errors we are simply concerned with a particular class of normal populations, and as such it can be treated by the more perfect concepts and terminology of present-day statistics. Accordingly, we shall here present the theory of errors from the points of view already developed in the preceding chapters and thus be able to deduce the results of the old theory with greater ease and clearer reasoning. We have, in fact, already discussed various problems connected with the normal population, and our main task will perhaps be recounting the same in the present context.

Consider the measurement of a physical quantity by means of an experimental process which may be more or less elaborate. Now it is a matter of common experience with an experimenter that if repeated observations are taken under conditions as uniform as possible, the measured values do not all coincide but instead fluctuate at random, the fluctuations being big or small depending on the accuracy of the measuring process. Clearly then, the measured values also differ at random from the true value (unknown) of the quantity, thereby committing errors which are called *experimental errors* or *errors of observation*. These errors are uncontrollable and random in nature, and hence are also called *random errors* or *accidental errors*.^{*} The experimental errors are presumably caused by numerous subtle and

^{*}Errors of another kind known as *systematic errors* may also be present which are more or less constant in nature arising, for example, from faulty calibration of the instruments, personal equation of the observer etc. These can, however, be eliminated or minimised to a great extent by care and caution on the part of the observer, but, at any rate, cannot be treated by means of a mathematical theory.

uncontrollable factors which again vary at random from one observation to another, e.g. fluctuations of temperature, pressure etc. of the surrounding, atmospherical disturbances for astronomical observations, undetectable vibrations of the instruments, reading a scale by way of eye-estimation and many other known and unknown factors.

Let X denote the measured value of the physical quantity whose hypothetical true value is m . Mathematically speaking, X is a random variable, and if we assume for a theoretical model that a measurement may yield any real number, X can be taken to be a continuous random variable having the spectrum $(-\infty, \infty)$. The probability density function of X should then naturally contain the true value m as a parameter.

The random variable

$$E = X - m \quad (18.1.1)$$

will be called the *error* in the measurement.

Let n measurements of the quantity, performed under uniform conditions, give the set of values : x_1, x_2, \dots, x_n which is a random sample of size n from the population of X . The corresponding errors are given by

$$e_i = x_i - m \quad (i = 1, 2, \dots, n) \quad (18.1.2)$$

so that e_1, e_2, \dots, e_n is a sample of size n from the population of E .

18.2 THE NORMAL LAW

Our first problem will be to determine the probability distribution of X , and we are going to show that it can be taken to be a normal distribution with mean m , the true value of the quantity and standard deviation σ which gives an inverse measure of precision of the measuring process.

It seems quite plausible to require at the outset that the true value m should be the best-fitting point to the distribution of the measured value X in some sense or other. If, in particular, we adopt the principle of least squares, we shall get $m = E(X)$ (cf. Remark Sec. 8.13). Thus it is reasonable to take m as the mean of the distribution of X , and hence the mean of the error variable E is zero.

HYPOTHESIS OF ELEMENTARY ERRORS. Since the error is believed to be produced by a large number of different and apparently unrelated causes, we may reasonably assume that each of these gives rise to an elementary error such that the elementary errors are mutually independent random variables, their sum being the total error E . Now we know that the Central Limit Theorem holds for many a sequence of random variables, and if the same is assumed hold for the elementary errors, then it follows that their sum E is approximately normally distributed. Since the mean of each elementary error is zero, the mean of E is also zero so that $X = E + m$ is approximately normal with mean m . The other parameter σ , the standard deviation of X , we know, gives an inverse measure of concentration of the probability mass in the distribution of X about the mean m and thus is an inverse measure of precision of measurement.

Formal derivation. The normal law of error may also be formally deduced from some simple starting hypotheses whose plausibility is guaranteed by common experience. The deduction will consist of two stages. In the first place, we shall show that a set of elementary postulates regarding the maximum likelihood estimate of the true value, \hat{m} leads to its unique determination, viz. $\hat{m} = \bar{x}$, the sample mean. Then, on the basis of this, we shall prove the normal law following a method originally due to Gauss.

POSTULATES FOR \hat{m}

1. \hat{m} is a simple function or, precisely, a continuously differentiable function of the sample values x_1, x_2, \dots, x_n .

2. \hat{m} is a symmetric function of x_1, x_2, \dots, x_n .

3. \hat{m} is independent of the origin of measurement, i.e. for any number h

$$\hat{m}(x_1 + h, x_2 + h, \dots, x_n + h) = \hat{m}(x_1, x_2, \dots, x_n) + h$$

4. m is independent of the unit of measurement, i.e. for any $k > 0$

$$\hat{m}(kx_1, kx_2, \dots, kx_n) = k\hat{m}(x_1, x_2, \dots, x_n)$$

or, in other words, \hat{m} is a homogeneous function of the first degree.

By Postulates 4 and 1

$$\begin{aligned} k\hat{m}(x_1, x_2, \dots, x_n) &= \hat{m}(kx_1, kx_2, \dots, kx_n) \\ &= \hat{m}(0, 0, \dots, 0) + k \sum x_i \left[\frac{\partial \hat{m}}{\partial x_i} \right]_{(\theta kx_1, \theta kx_2, \dots, \theta kx_n)} \quad (0 < \theta < 1) \end{aligned}$$

Making $k \rightarrow +0$

$$\hat{m}(0, 0, \dots, 0) = 0$$

and hence dividing by k and again making $k \rightarrow +0$, we have

$$\hat{m}(x_1, x_2, \dots, x_n) = \sum x_i \left[\frac{\partial \hat{m}}{\partial x_i} \right]_{(0, 0, \dots, 0)}$$

By Postulate 2 the constants $\left[\frac{\partial \hat{m}}{\partial x_i} \right]_{(0, 0, \dots, 0)}$ must all be the same, say, c . Hence

$$\hat{m}(x_1, x_2, \dots, x_n) = c \sum x_i$$

By Postulate 3 $c \sum (x_i + h) = c \sum x_i + h$ or $c = 1/n$. Hence $\hat{m} = \bar{x}$, the sample mean.

PROOF OF THE NORMAL LAW. We shall now make another hypothesis, viz. that the probability distribution of the error E is independent of the true value m . This hypothesis is also in keeping with experience, from which it follows that the density function of X is a function of $x - m$, i.e. we can write

$$f(x; m) = g(x - m)$$

The likelihood function of the sample is then given by

$$L = g(x_1 - m)g(x_2 - m) \dots g(x_n - m)$$

or

$$\log L = \sum \log g(x_i - m)$$

The likelihood equation is

$$\frac{\partial \log L}{\partial m} = 0 \quad \text{or} \quad \sum \frac{g'(x_i - m)}{g(x_i - m)} = 0$$

Setting, for convenience, $G(x-m) = \frac{g'(x-m)}{g(x-m)}$ the above equation reduces to

$$\sum G(x_i - m) = 0 \quad (i)$$

Since this would lead to the solution $m = \bar{x}$ or

$$\sum (x_i - m) = 0 \quad (ii)$$

we have

$$\sum \{G'(x_i - m) + \lambda\} d(x_i - m) = 0$$

where λ is a Lagrange's multiplier. It follows that

$$G'(x_i - m) + \lambda = 0 \quad (i = 1, 2, \dots, n)$$

or, in fact,

$$G'(x - m) + \lambda = 0$$

So

$$G(x - m) + \lambda(x - m) + \mu = 0 \quad [\mu, \text{ a constant}]$$

By (i) and (ii) $\mu = 0$ so that

$$\frac{g'(x-m)}{g(x-m)} + \lambda(x-m) = 0$$

Integrating we get

$$f(x; m) = g(x - m) = A e^{-\frac{1}{2}\lambda(x-m)^2} \quad [A, \text{ a constant}]$$

Now we must have

$$\int_{-\infty}^{\infty} f(x; m) dx = 1 \quad (iii)$$

For the convergence of this infinite integral λ must be positive, and hence writing $\lambda = 1/\sigma^2$ we have from (iii) $A = 1/\sqrt{2\pi\sigma}$. So

$$f(x; m) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-m)^2/2\sigma^2}$$

i.e. X is normal (m, σ). This proves the normal law of error which is also known as the *Gaussian law* in the theory of errors.

18.3 SOME DEFINITIONS

The theory of errors has a traditional terminology of its own. Thus, in the theory of errors the maximum likelihood estimate of the true value, \hat{m} is called the *most probable value* of the quantity, the standard deviation σ the *root mean square error* or simply the *mean square error* of measurement, i.e. of X .

Modulus of precision. We have seen that σ gives an inverse measure of precision of measurement. To get a direct measure, set

$$h = \frac{1}{\sqrt{2}\sigma} \quad (18.3.1)$$

h will be called the *modulus of precision* of measurement or of X , which provides the required direct measure of precision. The normal law then assumes the form

$$f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2(x-m)^2} \quad (18.3.2)$$

Error function. The random variable E is obviously normal $(0, \sigma)$, and hence, writing in terms of h , we have

$$P(|E| < e) = \frac{2h}{\sqrt{\pi}} \int_0^e e^{-h^2 e^2} de = \frac{2}{\sqrt{\pi}} \int_0^{he} e^{-x^2} dx = \Theta(he)$$

where

$$\Theta(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx \quad (18.3.3)$$

is called the *error function* which is also sometimes denoted by $\text{erf}(x)$. Tables were prepared for this error function which presumably served as a substitute for the now commonly used tables of standard normal distribution function. The relation between the two may be easily verified to be the following :

$$\Theta(x) = 2\Phi(\sqrt{2}x) - 1 \quad (18.3.4)$$

Probable error. If $\pm Q$ denote the quartiles of the error E , then Q is called the *probable error* of X . Q is obtained in terms of σ as follows. We have

$$P(E > Q) = .25$$

or

$$P\left(\frac{E}{\sigma} > \frac{Q}{\sigma}\right) = .25$$

Since E/σ is standard normal, we have from Table I $Q/\sigma = 0.67$, or from more accurate tables $Q/\sigma = 0.6745$, i.e.

$$Q = 0.6745 \sigma \quad (18.3.5)$$

The probable error gives another inverse measure of precision of the measuring process.

18.4 ESTIMATION

Maximum likelihood method. We know from Sec. 14.2 that $\hat{m} = \bar{x}$, $\hat{\sigma} = S$. We know further that a better estimate of σ^2 , particularly for small samples, is s^2 , where s^2 is the unbiased estimate of the population variance; s can also be obtained as a maximum likelihood estimate of σ if, instead of the parent population, we consider a sample of unit size from the population of the statistic s^2 (or S^2) (cf. Remark 1 Sec. 14.2), and let us write

$$\sigma^\dagger = s \quad (18.4.1)$$

Least square method. Since the measured values are attempted approximations to the true value m , we may also employ the principle of least squares as a method of estimation of m , i.e. we fit a point to the observed points x_1, x_2, \dots, x_n by the least square principle, and take the best-fitting point to be an estimate of m . For this we have to minimise

$$\sum (x_i - m)^2 = \sum e_i^2 \quad (18.4.2)$$

as a function of m (cf. Remark Sec. 8.13). This is the simplest form of the principle of least squares which states that the sum of the squares of the errors should be a minimum. The normal equation is

$$\sum (x_i - m^*) = 0 \quad (18.4.3)$$

which gives

$$m^* = \bar{x} \quad (18.4.4)$$

Thus the least square estimate of m coincides with its maximum likelihood estimate. The method of maximum likelihood, in fact, leads

to the principle of least squares, for the likelihood function of the sample is

$$L = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m)^2}$$

so that, for fixed σ , L is maximum when $\sum (x_i - m)^2$ is minimum, which is the principle of least squares.

Remark. The method of least squares has one advantage over that of maximum likelihood, viz. it does not require any knowledge regarding the population distribution. If, however, the population is assumed to be normal, then the method of maximum likelihood implies the principle of least squares, as we have just now seen.

The *residuals* of the sample are given by

$$v_i = x_i - m^* = x_i - \bar{x} \quad (18.4.5)$$

The normal equation (18.4.3) states that

$$\sum v_i = 0 \quad (18.4.6)$$

Also

$$\sum e_i = n(\bar{x} - m)$$

So

$$v_i = e_i - \frac{1}{n} \sum e_i = -\frac{e_1}{n} - \frac{e_2}{n} - \dots + \frac{(n-1)e_i}{n} - \dots - \frac{e_n}{n} \quad (i = 1, 2, \dots, n) \quad (18.4.7)$$

This equation in terms of the corresponding random variables will be

$$V_i = -\frac{E_1}{n} - \frac{E_2}{n} - \dots + \frac{(n-1)E_i}{n} - \dots - \frac{E_n}{n} \quad (18.4.8)$$

Since X_1, X_2, \dots, X_n are mutually independent random variables each normal (m, σ) , E_1, E_2, \dots, E_n are also mutually independent each normal $(0, \sigma)$, and hence V_i , which is a linear combination of E_1, E_2, \dots, E_n , is normally distributed such that

$$E(V_i) = 0$$

and

$$\text{var}(V_i) = \frac{n-1}{n^2} \sigma^2 + \frac{(n-1)^2}{n^2} \sigma^2 = \frac{n-1}{n} \sigma^2$$

i.e. each residual V_i is normal $\left(0, \sqrt{\frac{n-1}{n}} \sigma\right)$.

Remark. The residuals V_1, V_2, \dots, V_n are not mutually independent but are restricted by $\sum V_i = 0$.

Written in terms of the residuals

$$\hat{\sigma} = S = \sqrt{\frac{\sum v_i^2}{n}} \quad (18.4.9)$$

$$\hat{Q} = 0.6745 S = 0.6745 \sqrt{\frac{\sum v_i^2}{n}} \quad (18.4.10)$$

and

$$\sigma^\dagger = s = \sqrt{\frac{\sum v_i^2}{n-1}} \quad (18.4.11)$$

$$Q^\dagger = 0.6745 s = 0.6745 \sqrt{\frac{\sum v_i^2}{n-1}} \quad (18.4.12)$$

It is customary to reckon the precision of measurement either by an estimate of the mean square error or that of the probable error. The above formulas give estimates of the mean square error and the probable error for a single measurement; estimates (18.4.9) and (18.4.10) may be used for large samples, but for small samples (18.4.11) and (18.4.12) are preferable.

Let us now find the corresponding quantities for the most probable value. By Theorem I Sec. 13.4 the most probable value \bar{X} is normal $(m, \sigma/\sqrt{n})$ so that

$$\sigma(\bar{X}) = \sigma/\sqrt{n} \quad (18.4.13)$$

$$Q(\bar{X}) = 0.6745 \sigma/\sqrt{n} \quad (18.4.14)$$

and their estimates are given by

$$\hat{\sigma}(\bar{X}) = \frac{S}{\sqrt{n}} = \frac{\sqrt{\sum v_i^2}}{n} \quad (18.4.15)$$

$$\hat{Q}(\bar{X}) = 0.6745 \frac{S}{\sqrt{n}} = 0.6745 \frac{\sqrt{\sum v_i^2}}{n} \quad (18.4.16)$$

and

$$\sigma^\dagger(\bar{X}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum v_i^2}{n(n-1)}} \quad (18.4.17)$$

$$Q^\dagger(\bar{X}) = 0.6745 \frac{s}{\sqrt{n}} = 0.6745 \sqrt{\frac{\sum v_i^2}{n(n-1)}} \quad (18.4.18)$$

Confidence interval. In old practice the result for the true value used to be presented in the form :

most probable value \pm an estimate of its probable error (or sometimes its mean square error), i.e. using (18.4.18), as

$$\bar{x} \pm 0.6745 \frac{s}{\sqrt{n}} \quad (18.4.19)$$

without, however, any precise indication how to use the estimate of the probable error as a correction to \bar{x} .

In modern statistics the use of probable error (or the like) has been completely and very successfully replaced by the concept of confidence interval. The confidence intervals for the parameters of a normal population have already been studied in Sec. 14.5, and it will be easily recognised that (18.4.19) gives nothing but the 50% values of the approximate confidence limits (14.5.5) for the population mean m . The exact confidence limits are, however, given by (14.5.3) and (14.5.4).

Example. A measurement was repeated 5 times with the following results :

2.03, 2.08, 2.03, 2.01, 2.05

Find the most probable value of the quantity and the corresponding probable error. Find also 95% confidence limits for the true value.

Set $x = 2 + .01x'$.

x	x'	x'^2
2.03	3	9
2.08	8	64
2.03	3	9
2.01	1	1
2.05	5	25
Total	20	108

$$\bar{x}' = 4.0, \quad a_2' = 21.6$$

$$S'^2 = 5.6, \quad S' = 2.4$$

Hence

$$\bar{x} = 2.040, \quad S = .024$$

By (18.4.18)

$$Q(\bar{X}) = .008$$

For 95% confidence limits, t_e is given by $P(t > t_e) = .025$ corresponding to 4 degrees of freedom, whence from Table III $t_e = 2.776$. Thus 95% confidence limits for the true value are $2.040 \pm .033 = 2.007, 2.073$.

18.5 WEIGHTED MEASUREMENTS

Weights of measurements. Sometimes it so happens that measurements of the same quantity or of different quantities are done by different processes whose relative accuracies are known. Thus let x_1, x_2, \dots, x_n be the results of n independent measurements of the same quantity or of different quantities having moduli of precision h_1, h_2, \dots, h_n respectively such that

$$h_1^2 : h_2^2 : \dots : h_n^2 = w_1 : w_2 : \dots : w_n \quad (18.5.1)$$

where w_1, w_2, \dots, w_n are known positive numbers which are called the *weights* of the corresponding measurements or of the corresponding random variables X_1, X_2, \dots, X_n . Obviously, the weights are not exactly determinate, for all of them can be multiplied by any constant, i.e. if w_1, w_2, \dots, w_n are the weights of X_1, X_2, \dots, X_n , then cw_1, cw_2, \dots, cw_n may be also taken to be their weights, c being any constant. Hence we may set

$$h_i = \sqrt{w_i} h \quad (18.5.2)$$

where h is an unknown constant, or for the standard deviations

$$\sigma_i = \sigma / \sqrt{w_i} \quad (18.5.3)$$

where

$$\sigma = 1 / \sqrt{2h} \quad (18.5.4)$$

Presently we are interested in the repeated measurements of a single quantity whose true value is m ; in that case the random variables X_1, X_2, \dots, X_n are mutually independent, X_i being normal $(m, \sigma / \sqrt{w_i})$ ($i = 1, 2, \dots, n$).

In order to realise how the weights arise in practical problems, we take the following example. Let a quantity be independently measured 10 times by the same method corresponding to modulus of precision h , and the results be x_1, x_2, \dots, x_{10} . Suppose the arithmetic means

$$y_1 = (x_1 + x_2)/2, y_2 = (x_3 + x_4 + x_5)/3, y_3 = (x_6 + \dots + x_{10})/5$$

are calculated, and the results are presented in the condensed form y_1, y_2, y_3 ; then the random variables Y_1, Y_2, Y_3 are normal each with mean m , the true value but moduli of precision $h_1 = \sqrt{2}h$,

$h_2 = \sqrt{3}h$, $h_3 = \sqrt{5}h$ respectively so that $h_1^2 : h_2^2 : h_3^2 = 2 : 3 : 5$. Hence y_1, y_2, y_3 may be taken to be the measured values of the given quantity corresponding to weights 2, 3, 5 respectively. This example illustrates an important method by which weights are determined in practice. The weights may, however, be assigned by other methods as well in other situations.

Modified empirical distribution. We note that here the measured values x_1, x_2, \dots, x_n do not form a sample from the same population, but, in fact, are n independent samples of unit size from n different normal populations having the same mean m but different standard deviations $\sigma_i = \sigma / \sqrt{w_i}$ ($i = 1, 2, \dots, n$). In order to represent the distribution of these measured values taking their weights into consideration, we construct a hypothetical discrete probability distribution, in which the spectrum consists of the points x_1, x_2, \dots, x_n and the total mass 1 is distributed to these points in the proportion of their weights, i.e. the mass at x_i is $w_i / \sum w_i$ ($i = 1, 2, \dots, n$). This is the modified empirical distribution for a set of weighted measurements. Let the characteristics of this empirical distribution be denoted by the corresponding notations for the characteristics of a sample (Sec. 12.4) with an overhead bar, so that the mean, variance etc. of this distribution are given by

$$\bar{x} = \frac{1}{W} \sum w_i x_i \quad (18.5.5)$$

which is the weighted arithmetic mean of x_1, x_2, \dots, x_n ,

$$\bar{s}^2 = \frac{1}{W} \sum w_i (x_i - \bar{x})^2$$

etc. where

$$W = \sum w_i \quad (18.5.6)$$

The corresponding random variables are

$$\bar{X} = \frac{1}{W} \sum w_i X_i \quad (18.5.7)$$

$$\bar{S}^2 = \frac{1}{W} \sum w_i (X_i - \bar{X})^2$$

etc. Since X_i is normal ($m, \sigma/\sqrt{w_i}$) and X_i 's are mutually independent, \bar{X} , a linear combination of X_i 's, is normal ($m, \sigma/\sqrt{W}$), or, in other words, we may regard \bar{X} as a measured value of the given quantity of weight W . It follows that the statistic $U = \sqrt{W}(\bar{X} - m)/\sigma$ is standard normal.

We may write

$$\bar{S}^2 = \frac{1}{W} \sum w_i (X_i - m)^2 - (\bar{X} - m)^2$$

which gives

$$E(\bar{S}^2) = (n-1)\sigma^2/W \quad (18.5.8)$$

This shows on putting

$$v \bar{S}^2 = n \bar{S}^2 \quad (v = n-1) \quad (18.5.9)$$

that $W \bar{S}^2/n$ is an unbiased estimate of σ^2 .

Also setting

$$Y_i = \sqrt{w_i}(X_i - m)/\sigma \quad (i = 1, 2, \dots, n)$$

we get from the above expression for \bar{S}^2 that

$$W \bar{S}^2/\sigma^2 = \sum Y_i^2 - U^2$$

where Y_i 's are mutually independent standard normal variates and $U = \sum \sqrt{w_i} Y_i / \sqrt{W}$ is a linear function of these such that the sum of the squares of the coefficients is unity, and hence by Theorem III Sec. 9.1 $\chi^2 = W \bar{S}^2/\sigma^2$ is χ^2 -distributed with $v = n-1$ degrees of freedom, and χ^2 and U or \bar{X} and \bar{S}^2 are independent.

Now χ^2 and U are independent, the former being χ^2 -distributed with $v = n-1$ degrees of freedom and the latter standard normal, so that by Theorem I Sec. 9.2 the statistic $t = \sqrt{n}(\bar{X} - m)/\bar{S}$ has a t -distribution with $v = n-1$ degrees of freedom.

Estimation

MAXIMUM LIKELIHOOD METHOD. Here X_i is normal ($m, \sigma/\sqrt{w_i}$), its density function being

$$f_{x_i}(x_i; m, \sigma) = \frac{\sqrt{w_i}}{\sqrt{2\pi}\sigma} e^{-\frac{w_i(x_i - m)^2}{2\sigma^2}} \quad (i = 1, 2, \dots, n)$$

Hence the joint likelihood function of x_1, x_2, \dots, x_n is given by

$$L = (2\pi)^{-n/2} \sigma^{-n} \sqrt{w_1 w_2 \dots w_n} e^{-\frac{1}{2\sigma^2} \sum w_i (x_i - m)^2}$$

We remember that the weights w_1, w_2, \dots, w_n are known, and m, σ are the only parameters to be estimated. Hence

$$\log L = -n \log \sigma - \frac{1}{2\sigma^2} \sum w_i (x_i - m)^2 + \text{const.}$$

and the likelihood equations are

$$\frac{\partial \log L}{\partial m} = 0, \quad \frac{\partial \log L}{\partial \sigma} = 0$$

The first equation gives

$$\sum w_i (x_i - m) = 0$$

or

$$\hat{m} = \bar{x} \quad (18.5.10)$$

and the second

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum w_i (x_i - m)^2 = 0$$

or

$$\hat{\sigma}^2 = W \bar{S}^2 / n \quad (18.5.11)$$

It follows from (18.5.8) that the estimate $\hat{\sigma}^2$ is not unbiased, which is only expected.

If σ_i and Q_i respectively denote the standard deviation and probable error of X_i , then

$$\hat{\sigma}_i = \hat{\sigma} / \sqrt{w_i} = \sqrt{W/nw_i} \bar{S} \quad (18.5.12)$$

$$\hat{Q}_i = 0.6745 \sqrt{W/nw_i} \bar{S}$$

Also for the most probable value we have

$$\hat{\sigma}(\bar{X}) = \hat{\sigma} / \sqrt{W} = \bar{S} / \sqrt{n} \quad (18.5.13)$$

$$\hat{Q}(\bar{X}) = 0.6745 \bar{S} / \sqrt{n}$$

If, however, we consider the likelihood function for a sample of unit size from the population of \bar{s}^2 (or \bar{S}^2), the maximum likeli-

hood estimate of σ^2 turns out to be the unbiased estimate $W\bar{s}^2/n$ which will be denoted by $\sigma^{\dagger 2}$, i.e.

$$\sigma^{\dagger 2} = W\bar{s}^2/n \quad (18.5.14)$$

and correspondingly we have

$$\sigma_i^{\dagger} = \sqrt{W/nw_i} \bar{s}, \quad Q_i^{\dagger} = 0.6745 \sqrt{W/nw_i} \bar{s} \quad (18.5.15)$$

$$\sigma^{\dagger}(\bar{X}) = \bar{s}/\sqrt{n}, \quad Q^{\dagger}(\bar{X}) = 0.6745 \bar{s}/\sqrt{n} \quad (18.5.16)$$

LEAST SQUARE METHOD. In this case the principle of least squares will consist in fitting a point to the modified empirical distribution of the measured values defined earlier, i.e. minimising $\sum w_i(x_i - m)^2/W$ or

$$\sum w_i(x_i - m)^2 = \sum w e_i^2 \quad (18.5.17)$$

—the weighted sum of the squares of the errors. This is the modified form of the principle of least squares. The normal equation on putting $m = m^*$ becomes

$$\sum w_i(x_i - m^*) = 0 \quad (18.5.18)$$

giving

$$m^* = \bar{x} \quad (18.5.19)$$

which is the same result as before. Here also the principle of least squares follows as a consequence of the principle of maximum likelihood; for fixed σ , a maximum of the likelihood function L clearly corresponds to a minimum of (18.5.17).

The residuals v_i are defined by

$$v_i = x_i - m^* = x_i - \bar{x} \quad (18.5.20)$$

From (18.5.18)

$$\sum w v_i = 0 \quad (18.5.21)$$

i.e. the weighted sum of the residuals is zero.

We have

$$\bar{S}^2 = \frac{1}{W} \sum w_i v_i^2 \quad (18.5.22)$$

and the formulas (18.5.12)–(18.5.16) may be easily written down in terms of the residuals. In particular, we have

$$\hat{Q}(\bar{X}) = 0.6745 \sqrt{\frac{\sum w_i v_i^2}{nW}} \quad (18.5.23)$$

$$Q^\dagger(\bar{X}) = 0.6745 \sqrt{\frac{\sum w_i v_i^2}{(n-1)W}} \quad (18.5.24)$$

CONFIDENCE INTERVAL. Considering the statistic $t = \sqrt{n}(\bar{x} - m)/\bar{s}$ whose sampling distribution is t -distributed with $\nu = n - 1$ degrees of freedom, a confidence interval for m having confidence coefficient $1 - \epsilon$ is easily found to be

$$\left(\bar{x} - \frac{\bar{s}t_\epsilon}{\sqrt{n}}, \bar{x} + \frac{\bar{s}t_\epsilon}{\sqrt{n}} \right) \quad (18.5.25)$$

where

$$P(t > t_\epsilon) = \frac{1}{2}\epsilon \quad (18.5.26)$$

Example. Work out the example of the previous section, assuming that the measurements have weights 1, 1, 3, 2, 3 respectively.

Set $x = 2 + .01x'$.

x	w	x'	wx'	wx'^2
2.03	1	3	3	9
2.08	1	8	8	64
2.03	3	3	9	27
2.01	2	1	2	2
2.05	3	5	15	75
Total	10	—	37	177

$$\bar{x}' = 3.7, \quad \bar{a}_2' = 17.7$$

$$S'^2 = 4.01, \quad S' = 2.0$$

so that

$$\bar{x} = 2.037, \quad \bar{S} = .020$$

By (18.5.24)

$$Q^\dagger(\bar{X}) = .007$$

As before $t_\epsilon = 2.776$, and hence by (18.5.25) the required confidence limits are $2.037 \pm .028 = 2.009, 2.065$.

18.6 INDIRECT OBSERVATIONS

Suppose we have to determine the values of a number of physical quantities whose true values are q_1, q_2, \dots, q_k , and the situation is such that these quantities cannot be directly measured but instead $n (\geq k)$ other quantities can be measured, whose true values m_1, m_2, \dots, m_n are known functions of q_1, q_2, \dots, q_k , and let x_1, x_2, \dots, x_n denote their independently measured values. Our problem then will be to find estimates of q_1, q_2, \dots, q_k on the basis of the observed values x_1, x_2, \dots, x_n . Here, for simplicity, we shall consider the case of linear functions only given by

$$\begin{aligned} m_1 &= a_1 q_1 + b_1 q_2 + \dots + f_1 q_k \\ m_2 &= a_2 q_1 + b_2 q_2 + \dots + f_2 q_k \\ &\dots \dots \dots \dots \dots \dots \dots \dots \\ m_n &= a_n q_1 + b_n q_2 + \dots + f_n q_k \end{aligned} \quad (18.6.1)$$

where the coefficients a 's, b 's etc. are known.

The errors in the measurements are

$$e_i = x_i - m_i = x_i - a_i q_1 - b_i q_2 - \dots - f_i q_k \quad (i = 1, 2, \dots, n) \quad (18.6.2)$$

Equally precise measurements. If σ is the standard deviation of each measurement, X_i is normal (m_i, σ), and hence the joint likelihood function of x_1, x_2, \dots, x_n is

$$L = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum (x_i - m_i)^2}$$

For fixed σ , maximising the likelihood function L amounts to minimising

$$\sum (x_i - m_i)^2 = \sum e_i^2 \quad (18.6.3)$$

which is the principle of least squares. This is, in fact, the extended form of the principle for combination of measurements of different quantities. We note that the unknown parameters in $\sum e_i^2$ are q_1, q_2, \dots, q_k , and the corresponding normal (or likelihood) equations are

$$\frac{\partial \sum e_i^2}{\partial q_1} = 0, \quad \frac{\partial \sum e_i^2}{\partial q_2} = 0, \dots, \frac{\partial \sum e_i^2}{\partial q_k} = 0$$

before, and the solution is obtained in the m_i^* 's which obviously satisfy the same constraint relations as m_i 's. This latter problem belongs to what is usually called the *Theory of Adjustment*.

Example 1. q_1 and q_2 are the true values of two quantities, and four other quantities whose true values m_1, m_2, m_3, m_4 are given by

$$m_1 = q_1 - q_2$$

$$m_2 = q_1 + 2q_2$$

$$m_3 = 5q_1 - 3q_2$$

$$m_4 = 3q_1 + q_2$$

are measured by independent and equally reliable experiments to be 2.5, 12.7, 19.2, 24.1 respectively. Find the least square estimates of q_1, q_2 as well as of m_1, m_2, m_3, m_4 .

For writing the normal equations (18.6.4) we prepare the following table :

a	b	x	a^2	b^2	ab	ax	bx
1	-1	2.5	1	1	-1	2.5	-2.5
1	2	12.7	1	4	2	12.7	25.4
5	-3	19.2	25	9	-15	96.0	-57.6
3	1	24.1	9	1	3	72.3	24.1
Total	—	—	36	15	-11	183.5	-10.6

The normal equations are

$$36q_1^* - 11q_2^* = 183.5$$

$$-11q_1^* + 15q_2^* = -10.6$$

Solving these we get

$$q_1^* = 6.291, \quad q_2^* = 3.907$$

Hence by (18.6.5)

$$m_1^* = 2.384, \quad m_2^* = 14.105, \quad m_3^* = 19.734, \quad m_4^* = 22.780$$

The final results may be written as

$$q_1^* = 6.29, \quad q_2^* = 3.91$$

$$m_1^* = 2.38, \quad m_2^* = 14.11, \quad m_3^* = 19.73, \quad m_4^* = 22.78$$

$$m_1^* = 2.392, m_2^* = 13.933, m_3^* = 19.654, m_4^* = 22.564$$

Thus the final results are

$$q_1^* = 6.24, \quad q_2^* = 3.85$$

$$m_1^* = 2.39, \quad m_2^* = 13.93, \quad m_3^* = 19.65, \quad m_4^* = 22.56$$

18.7 EXERCISES

1. A board, with a pair of rectangular axes marked on it, is placed on a horizontal table, and a shot is dropped upon it from a height being aimed at the origin. If (X, Y) denotes the random point of meeting of the shot with the board, then assuming that X and Y are independent and that the bivariate probability distribution of (X, Y) is symmetrical about the origin, prove that each of X, Y is normally distributed with zero mean.

2. If the random variables X_1, X_2, \dots, X_n denote measured values of n quantities whose true values are m_1, m_2, \dots, m_n with moduli of precision h_1, h_2, \dots, h_n respectively, then show that the linear combination $X = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$ is a measured value of a quantity whose true value is m with modulus of precision h given by

$$m = a_1 m_1 + a_2 m_2 + \dots + a_n m_n$$

$$\frac{1}{h^2} = \frac{a_1^2}{h_1^2} + \frac{a_2^2}{h_2^2} + \dots + \frac{a_n^2}{h_n^2}$$

Deduce that if w_1, w_2, \dots, w_n are the weights of X_1, X_2, \dots, X_n respectively, the weight w of X is given by

$$\frac{1}{w} = \frac{a_1^2}{w_1} + \frac{a_2^2}{w_2} + \dots + \frac{a_n^2}{w_n}$$

3. x_1, x_2, \dots, x_n are independent and equally precise measurements of a physical quantity. Show that the modulus of precision of any linear combination $c_1 x_1 + c_2 x_2 + \dots + c_n x_n$, where $c_1 + c_2 + \dots + c_n = 1$, is maximum when the linear combination happens to be the arithmetic mean of the measurements.

4. If x_1, x_2, \dots, x_n are the results of n independent measurements of a single quantity having weights w_1, w_2, \dots, w_n respectively, then show that the weight of the i th residual is $w_i W / (W + w_i)$ where $W = \sum w_i$.

5. The following are the results of 10 of the many observations of the outer diameter of Saturn's ring by Bessel with the heliometer at the Keonigsberg Observatory, the measured values being reduced to the mean distance of the Saturn from the Sun :

$$38''.91, \quad 39''.32, \quad 38''.93, \quad 39''.31, \quad 39''.17$$

$$39''.04, \quad 39''.57, \quad 39''.46, \quad 39''.30, \quad 39''.03$$

Assuming that the observations were made under similar conditions, find the most probable value of the quantity, estimates of the mean square error and probable error of a single measurement and those of the most probable value. Find also 50% and 95% confidence limits for the true value.

6. A physical quantity was measured by 6 different methods with the results : 0.690, 0.681, 0.673, 0.687, 0.677, 0.675, the weights of the measurements being 2, 1, 1, 1, 3, 4 respectively. Compute the most probable value of the quantity, estimates of the probable errors of the individual measurements and of the most probable value and 95% confidence limits for the true value.

7. The measured values of 4 quantities, whose true values m_1, m_2, m_3, m_4 are given by

$$m_1 = q_1 + q_2 - 2q_3$$

$$m_2 = q_1 + 2q_2 + 2q_3$$

$$m_3 = 2q_1 + 5q_2 - q_3$$

$$m_4 = 3q_1 - q_2 + 7q_3$$

where q_1, q_2, q_3 are the true values of three other quantities, are 4, 15, 26, 10 respectively. Find the best values of q_1, q_2, q_3 and m_1, m_2, m_3, m_4 according to the principle of least squares, if the measurements are (a) equally precise and (b) have weights 1, 1, 2, 4 respectively.

TABLES

If $F(x)$ denotes the distribution function of a random variable X , we define $\bar{F}(x)$ by

$$\bar{F}(x) = 1 - F(x) = P(X > x)$$

which, for a continuous distribution, represents the area of the tail of the density curve to the right of the point x .

Table I. Standard normal distribution

Here $\bar{\phi}(x)$ is tabulated as a function of x , where $\phi(x)$ denotes, in special, the standard normal distribution function. (Fig. 39)

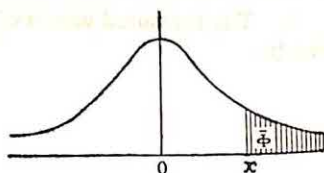


Fig. 39

Table II. χ^2 -distribution

The table gives the points χ^2 for different values of \bar{F} and the number of degrees of freedom n . (Fig. 40)

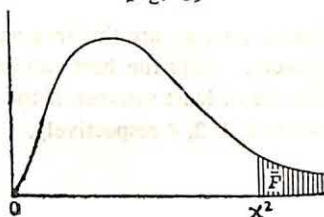


Fig. 40

Table III. t -distribution

The table gives the points t for different values of \bar{F} and the number of degrees of freedom n . (Fig. 41)

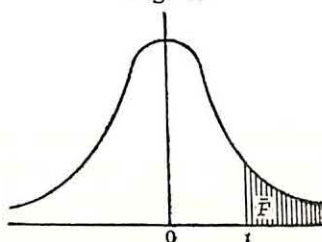


Fig. 41

Table IV. F -distribution

In Table IV, $\bar{F} = .05$, i.e. the 5% F -points are tabulated for different values of the parameters m and n . (Fig. 42)

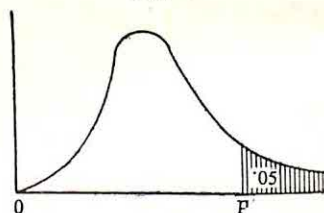


Fig. 42

Table V. F -distribution

In Table V, $\bar{F} = .01$, i.e. the 1% F -points are tabulated for different values of the parameters m and n . (Fig. 43)

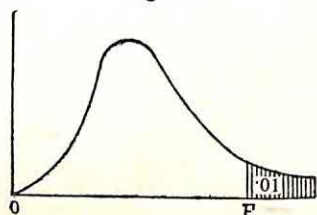


Fig. 43

Table I. Standard Normal Distribution

x	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09
0.0	.5000	.4960	.4920	.4880	.4841	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4091	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2644	.2611	.2579	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2207	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1563	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1094	.1075	.1057	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

The above table is abridged from Table II₁ of Fisher & Yates: "Statistical Tables for Biological, Agricultural and Medical Research" published by Oliver & Boyd Ltd., Edinburgh, and by permission of the authors and publishers.

Table II. χ^2 -Distribution

$\frac{\bar{F}}{n}$.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.000	.001	.004	.016	.064	.148	.455	1.074	1.642	2.706	3.841	5.4126	.635	10.827
2	.020	.040	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	.872	1.134	1.635	2.204	3.073	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697

16	5'812	6'614	7'962	9'312	11'152	12'624	15'338	18'418	20'465	23'542	26'296	29'633	32'000	39'252
17	6'408	7'255	8'672	10'085	12'002	13'531	16'338	19'511	21'615	24'769	27'587	30'995	33'409	40'790
18	7'015	7'906	9'390	10'865	12'857	14'440	17'338	20'601	22'760	25'989	28'869	32'346	34'805	42'312
19	7'633	8'567	10'117	11'651	13'716	15'352	18'338	21'689	23'900	27'204	30'144	33'687	36'191	43'820
20	8'260	9'237	10'851	12'443	14'578	16'266	19'337	22'775	25'038	28'412	31'410	35'020	37'566	45'315
21	8'897	9'915	11'591	13'240	15'445	17'182	20'337	23'858	26'171	29'615	32'671	36'343	38'932	46'797
22	9'542	10'600	12'338	14'041	16'314	18'101	21'337	24'939	27'301	30'813	33'924	37'659	40'289	48'268
23	10'196	11'293	13'091	14'848	17'187	19'021	22'337	26'018	28'429	32'007	35'172	38'968	41'638	49'728
24	10'856	11'992	13'848	15'659	18'062	19'943	23'337	27'096	29'553	33'196	36'415	40'270	42'980	51'179
25	11'524	12'697	14'611	16'473	18'940	20'867	24'337	28'172	30'675	34'382	37'652	41'566	44'314	52'620
26	12'198	13'409	15'379	17'292	19'820	21'792	25'336	29'246	31'795	35'563	38'885	42'856	45'642	54'052
27	12'879	14'125	16'151	18'114	20'703	22'719	26'336	30'319	32'912	36'741	40'113	44'140	46'963	55'476
28	13'565	14'847	16'928	18'939	21'588	23'647	27'336	31'391	34'027	37'916	41'337	45'419	48'278	56'893
29	14'256	15'574	17'708	19'768	22'475	24'577	28'336	32'461	35'139	39'087	42'557	46'693	49'588	58'302
30	14'953	16'306	18'493	20'599	23'364	25'508	29'336	33'530	36'250	40'256	43'773	47'962	50'892	59'703

The above table is abridged from Table IV of Fisher & Yates : "Statistical Tables for Biological, Agricultural and Medical Research" published by Oliver & Boyd Ltd., Edinburgh, and by permission of the authors and publishers.

Table III. *t*-Distribution

\bar{F} n	.45	.40	.35	.30	.25	.20	.15	.10	.05	.025	.01	.005	.0005
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073

16	.128	2.58	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.128	.257 ^a	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.27	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

The above table is reproduced from Table III of Fisher & Yates : "Statistical Tables for Biological, Agricultural and Medical Research," published by Oliver & Boyd Ltd., Edinburgh, and by permission of the authors and publishers.

Table IV. *F*-Distribution : 5 Percent Points

$\frac{m}{n}$	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.25	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

The above table is reproduced from Table V of Fisher & Yates :
 "Statistical Tables for Biological, Agricultural and Medical Research"
 published by Oliver & Boyd Ltd., Edinburgh, and by permission of the
 authors and publishers.

Table V. *F*-Distribution : 1 Percent Points

$n \backslash m$	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

The above table is reproduced from Table V of Fisher & Yates :
 "Statistical Tables for Biological, Agricultural and Medical Research"
 published by Oliver & Boyd Ltd., Edinburgh, and by permission of the
 authors and publishers.

ANSWERS AND HINTS

3.4

1. $P(\overline{A} + \overline{B}) = P(\overline{AB}) = 1 - P(AB)$; $P(\overline{A} \overline{B}) = P(\overline{A+B}) = 1 - P(A+B) = 1 - P(A) - P(B) + P(AB)$. Now $AB \overline{AB} = O$ and $AB + \overline{AB} = B$ so that $P(B) = P(AB) + P(\overline{AB})$; $P(\overline{A} + B) = P(\overline{A}) + P(B) - P(\overline{A}B) = 1 - P(A) + P(AB)$.

2. Required event $= (A - AB) + (B - AB)$.

3. $P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 A_2) \leq P(A_1) + P(A_2)$. Use induction.

4. $\binom{n}{r} / 2^n$

5. $1/2$

6. If A —at least one spade, then \overline{A} —none spade. $P(\overline{A}) = \binom{39}{2} / \binom{52}{2} = 19/34$; $P(A) = 15/34$.

7. A, B, C —both balls white, red, black respectively; required event is $A + B + C$ where A, B, C are pairwise exclusive. $n = 625$, $m(A) = 30$, $m(B) = 42$, $m(C) = 135$; $P(A + B + C) = 207/625$.

8. $2/9$

9. The probabilities in question are respectively $1 - (5/6)^4 \simeq .52$ and $1 - (35/36)^{24} \simeq .49$.

10. $(5/6)^k < 1/2$ or $k \geq 4$. Hence the answer is 4.

11. $2/n$

13. Let the $m+n$ tosses be represented by $m+n$ rooms. Total number of event points $= 2^{m+n}$. Since $m > n$, only one run of m heads is possible which begins at the 1st or 2nd... or $(n+1)$ th room. Except for the first and last cases, the rooms just before and after the run of m heads must be sealed off with tails in order that the run of heads may not get longer. In the first case the room just after and in the last case the room just before the run of heads must

be filled with a tail. Hence the event 'run of m heads' contains $(n-1)2^{n-2} + 2 \cdot 2^{n-1} = (n+3)2^{n-2}$ points, and the first result follows.

Let p_i denote the probability of exactly i consecutive leads. Then $p_m = (n+3)/2^{m+2}$, $p_{m+1} = (n+2)/2^{m+3}$, ..., $p_{m+n-2} = 5/2^{m+n}$ by the first part, but $p_{m+n-1} = 2/2^{m+n} = 1/2^{m+n-1}$ and $p_{m+n} = 1/2^{m+n}$; $p_m + p_{m+1} + \dots + p_{m+n} = [(n+3)2^{n+2} + (n+2)2^{n+1} + \dots + 5 \cdot 2^4 + 4 \cdot 2^3 + 3 \cdot 2^2 + 2 \cdot 2 + 1] - 1/2^{m+n+4}$. Differentiating the relation $1 + x + x^2 + \dots + x^{n+3} = (x^{n+4} - 1)/(x - 1)$ and putting $x = 2$, the second result follows.

$$15. \quad n = \binom{52}{2} + \binom{52}{4} + \dots + \binom{52}{52} = 2^{51} - 1, \quad m(A) = \binom{26}{1}^2 + \binom{26}{2}^2 + \dots + \binom{26}{26}^2 = \frac{52!}{(26!)^2} - 1 \text{ so that } P(A) = \left\{ \frac{52!}{(26!)^2} - 1 \right\} / (2^{51} - 1).$$

$$16. \quad 1 - \frac{\binom{48}{13}}{\binom{52}{13}} \approx 0.696.$$

$$17. \quad \frac{\binom{4}{4} \binom{48}{22}}{\binom{52}{26}} = \frac{46}{833}$$

18. If the person has to buy k tickets, then

$$\frac{1}{2} < 1 - \binom{10000-k}{100} / \binom{10000}{100} \approx 1 - (.99)^k$$

or $(.99)^k < .5$. This gives $k \geq 70$ so that the required answer is 70.

19. There are 13 face values each consisting of 4 cards. By

$$(3.1.14) \text{ the result} = 4^{13} / \binom{52}{13} \approx 0.000106.$$

$$20. \text{ By (3.1.14) the result} = \frac{\binom{3}{1} \binom{2}{1} \binom{4}{2}}{\binom{9}{4}} = 2/7.$$

21. Noting that the 4 suits can be arranged among themselves in $4!$ ways, the result, by (3.1.14), is $4! \binom{13}{5} \binom{13}{4} \binom{13}{3} \binom{13}{1} / \binom{52}{13} \approx 0.129$.

22. Let the r tickets drawn be placed in r rooms. The event $x_i = s$ means that the i th room is occupied by ticket no. s , so that rooms no. $1, 2, \dots, i-1$ are open to be filled by tickets no. $1, 2, \dots, s-1$ and rooms no. $i+1, \dots, r$ by tickets no. $s+1, \dots, n$. Hence etc.

23. If p_i denotes the probability given by (3.1.15) ($i \geq 1$) and $p_0 = N_1/N$, the result $= p_0 + p_2 + \dots$

$$24. \quad 3036/54145$$

25. Random experiment consists in drawing $k+1$ balls. Total number of event points $= (N_1 + N_2)(N_1 + N_2 - 1) \cdots (N_1 + N_2 - k)$. The required event means that the $(k+1)$ th drawing yields a white ball and hence contains $(N_1 + N_2 - 1)(N_1 + N_2 - 2) \cdots (N_1 + N_2 - k) N_1$ event points so that the answer is $N_1/(N_1 + N_2)$.

26. Continue to draw the remaining balls (which are of the same colour) and arrange all the balls drawn in N different rooms. The required event is equivalent to the event that the ball in the last room is white.

27. By (3.1.16) either probability $= 5/72$.

28. Let the balls be numbered $1, 2, \dots, r$. Total number of event points is n^r , since ball no. 1 may be placed in any one of the n cells and the same is true for balls no. 2, 3, \dots, r . If now i balls are placed in a given cell, we are left with $r-i$ balls to be distributed in $n-1$ cells, and since i balls can be chosen from r balls in $\binom{r}{i}$ ways, the

number of event points contained in the required event $= \binom{r}{i} (n-1)^{r-i}$.

Hence the first result follows. Now $p_i/p_{i+1} - 1 = n\{i+1 - (r+1)/n\}/(r-i)$ which leads to the second conclusion.

29. As in Ex. 28, the total number of event points $= (a+b)^n$. The number of ways in which $n-i$ objects can be distributed to a men and i objects to b women $= \binom{n}{i} a^{n-i} b^i$ so that the number of event points contained in the required event $= \binom{n}{1} a^{n-1} b + \binom{n}{3} a^{n-3} b^3 + \dots = \frac{1}{2}\{(a+b)^n - (a-b)^n\}$.

30. If a given cell contains i particles, we are left with $r-i$ particles to be placed in $n-1$ cells. Hence, by Ex. 13, P. 30, the number of event points contained in the required event $= \binom{n+r-i-2}{r-i}$. Hence etc.

31. See Ex. 14, P. 31. The total number of event points $= \binom{n}{r}$. If a given cell is empty, we have $n-1$ cells into which r particles are placed so that the required event contains $\binom{n-1}{r}$ points.

32. $1/3$

33. See Sec. 3.2 Ex. 5. Answers: $1/3, 3/8$

34. $\sum_{k=1}^{10} (-1)^{k-1}/k!$

35. A_1 —ball transferred is white, A_2 —ball transferred is black, X —ball drawn is white. Use (3.2.7). Answer: $5/21$

36. $61/140, 40/61$

37. $1/3$

38. Consider the first two urns only. Using (3.2.7) the probability of drawing a white ball from the second urn is $7/12$ and hence that of drawing a black ball is $5/12$. Similarly, considering the second and third urns, by (3.2.7) the required probability is $31/60$.

39. Let p_i —probability of drawing a white ball from the i th urn, so that the probability of drawing a black ball from the i th urn $= 1 - p_i$. Considering the i th and $(i+1)$ th urns and using (3.2.7)

$$p_{i+1} = \frac{N_i + 1}{N + 1} p_i + \frac{N_i}{N + 1} (1 - p_i) \quad (i = 1, 2, \dots, n-1)$$

But $p_1 = N_1/N$ which gives $p_2 = N_1/N$ and so on.

40. If A, B are independent, $P(AB) = P(A)P(B)$, and $A = AB + A\bar{B}$ where $AB\bar{A}\bar{B} = O$ so that $P(A) = P(AB) + P(A\bar{B}) = P(A)P(B) + P(A\bar{B})$, or $P(A\bar{B}) = P(A)\{1 - P(B)\} = P(A)P(\bar{B})$. This shows that A and \bar{B} are independent.

41. $P[A(B+C)] = P(AB+AC) = P(AB) + P(AC) - P(ABC) = P(A)P(B) + P(A)P(C) - P(A)P(BC) = P(A)[P(B) + P(C) - P(BC)] = P(A)P(B+C)$ which proves that A and $B+C$ are independent. By Ex. 40 \bar{A} and $\bar{B}+\bar{C} = \bar{B}\bar{C}$ are independent etc.

42. By (3.1.9) the required probability $= \Sigma p_1 - \Sigma p_1 p_2 + \Sigma p_1 p_2 p_3 - \dots = (1 - p_1)(1 - p_2)\dots(1 - p_n)$.

43. Not mutually independent.

4.9

1. A —spade, B —heart or diamond, C —queen. First Answer $= P\{A, B, C\} = P(A)P(B)P(C) = \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{13} = 1/104$. Second answer $= 3!/104 = 3/52$.

2. Answer = $1 - \frac{3}{8} \cdot \frac{2}{3} = \frac{3}{8}$.

3. (a) 5/16 (b) 1/2 (c) 13/16

4. 80/243

5. 1/64

6. $1 - (.8)^{10} - 2(.8)^9 \simeq .624$, 32

7. $(\frac{3}{8})^4$

8. 1568/6561

9. $p = 1/5$, $q = 5/6$. Answer = $\binom{n}{2} p^2 q^{n-2} + \binom{n}{4} p^4 q^{n-4} + \dots$
 $= \frac{1}{2} \{ (q+p)^n + (q-p)^n \}$.

10. (a) 16 or 17 (b) 33

12. Use (4.4.3).

13. See Ex. 3, Sec. 4.4.

14. $\frac{1}{2} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}$

15. $\binom{2n-i-1}{n-1} \left(\frac{1}{2}\right)^{2n-i}$

16. Let E_1, E_2, E_3 denote two trials of the experiment of observing if A, B, C respectively attends the class, i.e. two Bernoulli trials with probability of success $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}$ respectively. We are concerned with the joint performance of E_1, E_2, E_3 which are independent. Required probability

$$= \binom{1}{2}^2 \binom{2}{1} \frac{2}{3} \cdot \frac{1}{3} \cdot \left(\frac{1}{4}\right)^2 + \binom{1}{2}^2 \cdot \left(\frac{1}{3}\right)^2 \cdot \binom{2}{1} \frac{3}{4} \cdot \frac{1}{4} + \binom{2}{1} \frac{1}{2} \cdot \frac{1}{2} \cdot \binom{2}{3}^2 \cdot \left(\frac{1}{4}\right)^2$$

$$+ \binom{1}{2}^2 \cdot \left(\frac{2}{3}\right)^2 \cdot \binom{2}{1} \frac{3}{4} \cdot \frac{1}{4} + \binom{2}{1} \frac{1}{2} \cdot \frac{1}{2} \cdot \left(\frac{1}{3}\right)^2 \cdot \left(\frac{3}{4}\right)^2 + \binom{1}{2}^2 \cdot \binom{2}{1} \frac{2}{3} \cdot \frac{1}{3} \cdot \left(\frac{3}{4}\right)^2$$

$$+ \binom{2}{1} \frac{1}{2} \cdot \frac{1}{2} \cdot \binom{2}{1} \frac{2}{3} \cdot \frac{1}{3} \cdot \binom{2}{1} \frac{3}{4} \cdot \frac{1}{4} = \frac{1}{4}$$

17. See Ex. 6 Sec. 4.4. (a) $[i^n - (i-1)^n]/6^n$ (b) $[(7-i)^n - (6-i)^n]/6^n$

18. Let X - required event. Then \bar{X} - at least one of the n tickets does not appear in k drawings, and hence if A_i - ticket no. i does not appear in k drawings ($i = 1, 2, \dots, n$), $\bar{X} = \sum A_i$. By (3.1.9) and using

symmetry $P(\overline{X}) = nP(A_1) - \binom{n}{2}P(A_1A_2) + \dots$. Now the probability that ticket no. 1 does not appear in a single drawing $= \binom{n-1}{m} / \binom{n}{m} = \frac{n-m}{n}$, and since the drawings are independent, $P(A_1) = (n-m)^k/n^k$. Similarly, A_1A_2 - tickets no. 1 and 2 both do not appear in k drawings so that $P(A_1A_2) = \left(\frac{n-2}{m}\right)^k / \left(\frac{n}{m}\right)^k = \left(\frac{n-m}{n}\right)^k \left(\frac{n-m-1}{n-1}\right)^k$ etc.

19. $n = 500$, $p = 1/365$, $np = 500/365 = \mu$. By (4.4.8) answer $\simeq e^{-\mu}\mu = .348$.

20. $5^4 e^{-5}/4!$ (approx.)

21. $1/e$ (approx.)

22. Required probability $= \sum p_{k_1} p_{k_2} \dots p_{k_i} q_{i+1} \dots q_{k_n}$ where k_1, k_2, \dots, k_n are all different, each having one of the values $1, 2, \dots, n$.

23. By (4.5.1) answer $= \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{3}{7} + \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{4}{7} + \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{4}{7} = 29/70$.

24. By (4.6.3) answer $= 8! / 2! 3! 3! 6^3$.

25. $9/50$

26. When a pair of dice are thrown the probability of six, $p_1 = 5/36$ and that of seven, $p_2 = 1/6$. Then $q_1 = 1 - p_1 = 31/36$, $q_2 = 1 - p_2 = 5/6$. Let A_n - A wins in $2n+1$ throws, i.e. six appears at the $(2n+1)$ th throw but the 1st, 3rd, ..., $(2n-1)$ th throws result in 'not six' and the 2nd, 4th, ..., $2n$ th throws result in 'not seven' ($n=0, 1, 2, \dots$).

Then $X = \sum_{n=0}^{\infty} A_n$ is the required event, where A_0, A_1, \dots are pairwise exclusive events connected with an infinite sequence of independent throws of a pair of dice. Hence $P(X) = \sum_{n=0}^{\infty} P(A_n) = \sum_{n=0}^{\infty} q_1^n q_2^n p_1 = p_1(1 - q_1 q_2)^{-1} = 30/61$.

27. A_n - the first person wins in $3n+1$ throws, i.e. the first head appears at the $(3n+1)$ th throw ($n=0, 1, 2, \dots$). A_n 's are pairwise exclusive events connected with an infinite sequence of Bernoulli trials with probability of success $p = \frac{1}{2}$. Probability of the first person's winning $= P(\sum_{n=0}^{\infty} A_n) = \sum_{n=0}^{\infty} P(A_n) = \sum_{n=0}^{\infty} (1-p)^{3n} p = 4/7$. Similarly, the probabilities of win by the second and third persons are $2/7$ and $1/7$ respectively.

28. Let X - event of scoring n points. Then $X = A + B$ where A, B are mutually exclusive events defined by : A - head in the last throw scoring 1 point and scoring $n-1$ points in all throws except the last and B - tail in the last throw scoring 2 points and $n-2$ points in all throws except the last. $P(A) = p_{n-1} \cdot \frac{1}{2}$, $P(B) = p_{n-2} \cdot \frac{1}{2}$. Hence $p_n = P(X) = \frac{1}{2}(p_{n-1} + p_{n-2})$ or $2p_n = p_{n-1} + p_{n-2}$. We have $(-2)^n(p_n - p_{n-1}) = (-2)^{n-1}(p_{n-1} - p_{n-2}) = \dots = 4(p_2 - p_1)$. But $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$, so that $p_n - p_{n-1} = (-1)^n/2^n$. Replacing n by $n-1$, $n-2, \dots, 2$ and adding $p_n = \frac{1}{3}\{2 + (-1)^n/2^n\} \rightarrow 2/3$ as $n \rightarrow \infty$.

29. Let A_1, A_2 - $(n-1)$ th day is dry, wet respectively. X - n th day is dry. Given $P(X|A_1) = p$, $P(X|A_2) = p'$, $P(A_1) = u_{n-1}$, $P(X) = u_n$. Then $P(A_2) = 1 - P(A_1) = 1 - u_{n-1}$. By (3.2.7) $u_n = u_{n-1}p + (1 - u_{n-1})p'$ or $u_n - (p - p')u_{n-1} - p' = 0$. For $p = 3/4$, $p' = 1/4$, $u_n - \frac{1}{2}u_{n-1} - \frac{1}{4} = 0$ or $2^n(2u_n - 1) = 2^{n-1}(2u_{n-1} - 1) = \dots = 2(2u_1 - 1) = 2$, since $u_1 = 1$. Hence $u_n = \frac{1}{2} + 1/2^n$.

30. $\pi = (1, 0, 0)$. (i) By (4.8.6) answer $= \pi_1 p_{11} p_{12} p_{23} + \pi_1 p_{12} p_{22} p_{23} + \pi_1 p_{13} p_{32} p_{23} = 0$ (ii) answer $= p_{31}^{(2)} = 1/6$ (iii) answer $= \pi_2^{(5)} = 0$.

31. First prove by induction that

$$P^n = \frac{1}{2-p-q} \begin{pmatrix} 1-q & 1-p \\ 1-q & 1-p \end{pmatrix} + \frac{(p+q-1)^n}{2-p-q} \begin{pmatrix} 1-p & p-1 \\ q-1 & 1-q \end{pmatrix}$$

Hence probability distribution at the n th trial $= P^{n-1}$

$$\begin{aligned} &= \frac{1}{2-p-q} (1-q, 1-p) + \frac{(p+q-1)^{n-1}}{2-p-q} [\pi_1(1-p) + \pi_2(q-1), \\ &\pi_1(p-1) + \pi_2(1-q)] \\ &\rightarrow \frac{1}{2-p-q} (1-q, 1-p) \text{ as } n \rightarrow \infty, \text{ since } |p+q-1| < 1. \end{aligned}$$

5.10

2. $G(x) = \frac{1}{2h} \int_{-h}^h F(x+t) dt$. If $x < x'$, $F(x+t) \leq F(x'+t)$ so that, on

integration from $-h$ to h , $G(x) \leq G(x')$. For $x-h \leq t \leq x+h$, $F(x-h) \leq F(t) \leq F(x+h)$ which gives $F(x-h) \leq G(x) \leq F(x+h)$. Making $x \rightarrow \infty$ or $-\infty$, we get $G(\infty) = 1$, $G(-\infty) = 0$. $G(x)$ is continuous everywhere.

3. Use (5.5.3). Answers : $F(x) = i(n+1)/n(i+1)$ ($i \leq x < i+1$) ($i = 1, 2, \dots, n$) ; $(n-3)/4n$, $(n-5)/6n$

4. $A = F(-\infty) = 0$, $D = F(\infty) = 1$, $1/6 = P(X=0) = F(0) - F(-0) = C - B$, $2/3 = P(X > 1) = 1 - F(1) = 1 - C$. Hence $B = 1/6$, $C = 1/3$.

5. Spectrum $0, 1, 2, \dots, 6$; corresponding probabilities : $1/16$, $1/8$, $3/16$, $1/4$, $3/16$, $1/8$, $1/16$

6. (a) Binomial (5, .4) (b) Spectrum : $0, 1, 2, 3, 4$; corresponding probabilities : $1/42$, $5/21$, $10/21$, $5/21$, $1/42$

$$7. x_i = i \ (i = 0, 1, \dots, n), f_i = \binom{2n-i}{n} \left(\frac{1}{2}\right)^{2n-i+1}$$

$$8. x_i = i \ (i = 0, 1, \dots), f_i = q^i p \ (q = 1 - p)$$

9. See Ex. 12 Sec. 4.9.

$$10. \frac{1}{n!} \int_{\mu}^{\infty} e^{-x} x^n dx = \frac{e^{-\mu} \mu^n}{n!} + \frac{1}{(n-1)!} \int_{\mu}^{\infty} e^{-x} x^{n-1} dx \text{ etc.}$$

11. See Poisson process, Sec. 5.6. The number of wars in t years is Poisson distributed with parameter $\mu = \lambda t$ where $\lambda = 1/15$, $t = 25$ so that $\mu = 5/3$, and the required probability is $e^{-5/3}$.

12. The number of misprints in t pages is a Poisson variate with parameter $\mu = \lambda t$ where $\lambda = 500/500 = 1$, $t = 1$ so that $\mu = 1$ and so the answer $= e^{-1} \left(1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!}\right) = 8/3e$.

13. The number of bacteria in t c.c. of water is Poisson distributed with parameter $\mu = \lambda t$ where $\lambda = 10^6/10^5 = 10$, $t = 1$ so that $\mu = 10$. Hence the required probability $= e^{-10}$.

14. $f(x) \geq 0$ everywhere and $\int_{-\infty}^{\infty} f(x) dx = \int_{-1}^1 |x| dx = 1$. Hence $f(x)$ is possible density function. $F(x) = 0$ ($x \leq -1$), $(1-x^2)/2$ ($-1 < x \leq 0$), $(1+x^2)/2$ ($0 < x \leq 1$), 1 ($x > 1$).

$$15. 2, 3/4$$

$$16. 1/\pi, 2 \tan^{-1} e/\pi, 2 - 4 \tan^{-1} e/\pi$$

17. Probability masses $1/3$, $1/6$, $1/6$, $1/3$ at $0, 1, 2, 3$ respectively

$$18. 1/2$$

$$20. 1/\sqrt{2}$$

21. $k/(1+k)$
 23. Uniform in $(0, 1)$. Calculate $F(x)$ for $|X|$.
 24. $X = \cos Y$, Y uniformly distributed over $(0, \pi)$. Use (5.9.3).
 25. $Y = \mu + \lambda \cot X$, X uniformly distributed over $(0, \pi)$. Use (5.9.3).
 26. e^{-y} ($0 < y < \infty$)
 27. $f(y) = e^{-(\log y)^{1/2}}/y \sqrt{2\pi}$ ($0 < y < \infty$)
 28. $2e^{-y^2} y^{2l-1}/\Gamma(l)$ ($0 < y < \infty$)
 30. $x_i = i^2$ ($0, 1, 2, \dots$); $f_i = e^{-\mu} \mu^i / i!$. Use (5.9.5).

6.8

1. Let X —no. of the ball, Y —colour no.; $x_i = i$ ($i = 0, 1, \dots, 8$), $y_j = j$ ($j = 0, 1, 2$). Spectrum of (X, Y) is given by $(x_i, y_j) = (i, j)$ ($i = 0, 1, \dots, 8$; $j = 0, 1, 2$); $f_{ij} = 1/9$ for $(i = 0, 1, 2, 3; j = 0)$, $(i = 4, 5, 6; j = 1)$, $(i = 7, 8; j = 2)$ and 0 for the rest of the values of i, j . $f_{x_i} = f_i = 1/9$ for all i ; $f_{y_j} = f_j = 4/9, 1/3, 2/9$ for $j = 0, 1, 2$ respectively. The required conditional distribution: $f_{i|0} = f_{i0}/f_{y_0} = 1/4$ for all i .

2. Let X —ball no., Y —colour no. $x_i = i$ ($i = 0, 1, 2, 3$), $y_j = j$ ($j = 0, 1, 2$). Spectrum of (X, Y) : $(x_i, y_j) = (i, j)$ ($i = 0, 1, 2, 3$; $j = 0, 1, 2$). $f_{ij} = 1/12$ for all i, j ; $f_i = 1/4$ for all i , $f_j = 1/3$ for all j so that $f_{ij} = f_i \cdot f_j$ for all i, j . Hence etc.

3. $(x_i, y_j) = (i, j)$ ($i = 1, 2, \dots, 13$; $j = 1, 2, 3, 4$). $f_{ij} = 1/52$ for all i, j .

4. $(x_i, y_j) = (i = 0, 1, 2, \dots; j = 0, 1, 2, \dots, n)$; $f_{ij} = e^{-\mu} \frac{\mu^i}{i!} \binom{n}{j} p^j (1-p)^{n-j}$

5. $f_x(x) = \sin x$ ($0 < x < \pi/2$), $f_y(y) = \sin y$ ($0 < y < \pi/2$)

6. By (6.3.6), $1 = K \int_{x=0}^1 \int_{y=0}^x xy dx dy = K/8$ so that $K = 8$. $f_x(x) =$

$4x^3$ ($0 < x < 1$), $f_y(y) = 4y(1-y^2)$ ($0 < y < 1$).

7. $3/8, 5/24, 5/8, 3/5$. Draw diagrams.

8. $(3x^2 - 8xy + 6y^2)/(6y^2 - 4y + 1)$ ($0 < x < 1, 0 < y < 1$),
 $(3x^2 - 8xy + 6y^2)/(3x^2 - 4x + 2)$ ($0 < x < 1, 0 < y < 1$)

9. $2x$ ($0 < x < 1$), $2(1-y)$ ($0 < y < 1$); $1/(1-y)$ ($y < x < 1$),
 $1/x$ ($0 < y < x$); $1/2$

10. $f_x(x|y) = |y|e^{-y^2x^2}/\sqrt{\pi}$, $f_y(y) = \lambda e^{-\lambda^2y^2}/\sqrt{\pi}$. By (6.5.8) $f(x, y) = \lambda|y|e^{-y^2(x^2+\lambda^2)}/\pi$ so that $f_x(x) = \lambda\pi^{-1} \int_{-\infty}^{\infty} |y|e^{-y^2(x^2+\lambda^2)} dy = 2\lambda\pi^{-1} \times \int_0^{\infty} y e^{-y^2(x^2+\lambda^2)} dy = \lambda/\pi(x^2 + \lambda^2)$.

11. Let X, Y denote the random points. (X, Y) is uniformly distributed over the unit square $R: 0 < x < 1, 0 < y < 1$. The required event is $|X - Y| < k$ or $Y > X - k$ if $X > Y$ and $Y < X + k$ if $Y > X$, i.e. (X, Y) lies in the part R' of R lying between $y = x \pm k$. Hence $R' = 1 - (1 - k)^2 = k(2 - k)$ so that by (6.4.2) the required probability $= R'/R = k(2 - k)$.

12. (X, Y) is uniformly distributed over the unit square $R: 0 < x < 1, 0 < y < 1$ and the event in question is $XY < k$, i.e. (X, Y) lies in $R': xy < k$. Hence $R' = k + k \int_k^1 dx/x = k(1 - \log k)$ etc.

13. (p, q) is uniformly distributed over $R: -1 < p < 1, -1 < q < 1$, and the required event is $p^2 \geq q$, i.e. (p, q) lies in $R': p^2 \geq q$. Hence $R = 4$, $R' = 2 + 2 \int_0^1 p^2 dp = 8/3$, and the answer $= 2/3$.

14. For $0 < x < 1$, $f_x(x) = \int_{-(1-x)}^{1-x} \frac{1}{2} dy = 1 - x$, and for $-1 < x < 0$, $f_x(x) = \int_{-1-x}^{1+x} \frac{1}{2} dy = 1 + x$. Similarly for $0 < y < 1$, $f_y(y) = 1 - y$ and for $-1 < y < 0$, $f_y(y) = 1 + y$.

15. $2\sqrt{a^2 - x^2}/\pi a^2$ ($-a < x < a$), $2\sqrt{a^2 - y^2}/\pi a^2$ ($-a < y < a$), $1/2\sqrt{a^2 - x^2}$ ($-\sqrt{a^2 - x^2} < y < \sqrt{a^2 - x^2}$)

16. Choosing A as the origin, let X, Y be the co-ordinates of P, Q respectively. X, Y are independent, (X, Y) uniformly distributed over the rectangle $R: 0 < x < a, a < y < a + b$. Required event: $X < (a + b)/2, Y > (a + b)/2$ and $Y < X + (a + b)/2$, i.e. (X, Y) lies in the

21. $k/(1+k)$
23. Uniform in $(0, 1)$. Calculate $F(x)$ for $|X|$.
24. $X = \cos Y$, Y uniformly distributed over $(0, \pi)$. Use (5.9.3).
25. $Y = \mu + \lambda \cot X$, X uniformly distributed over $(0, \pi)$. Use (5.9.3).
26. e^{-y} ($0 < y < \infty$)
27. $f(y) = e^{-(\log y)^{2/2}}/y \sqrt{2\pi}$ ($0 < y < \infty$)
28. $2e^{-y^2} y^{2l-1}/\Gamma(l)$ ($0 < y < \infty$)
30. $x_i = i^2$ ($0, 1, 2, \dots$); $f_i = e^{-\mu^i}/i!$. Use (5.9.5).

6.8

1. Let X —no. of the ball, Y —colour no.; $x_i = i$ ($i = 0, 1, \dots, 8$), $y_j = j$ ($j = 0, 1, 2$). Spectrum of (X, Y) is given by $(x_i, y_j) = (i, j)$ ($i = 0, 1, \dots, 8$; $j = 0, 1, 2$); $f_{ij} = 1/9$ for $(i = 0, 1, 2, 3; j = 0)$, $(i = 4, 5, 6; j = 1)$, $(i = 7, 8; j = 2)$ and 0 for the rest of the values of i, j . $f_{x_i} = f_{i.} = 1/9$ for all i ; $f_{y_j} = f_{.j} = 4/9, 1/3, 2/9$ for $j = 0, 1, 2$ respectively. The required conditional distribution: $f_{i|0} = f_{i0}/f_{y_0} = 1/4$ for all i .

2. Let X —ball no., Y —colour no. $x_i = i$ ($i = 0, 1, 2, 3$), $y_j = j$ ($j = 0, 1, 2$). Spectrum of (X, Y) : $(x_i, y_j) = (i, j)$ ($i = 0, 1, 2, 3$; $j = 0, 1, 2$). $f_{ij} = 1/12$ for all i, j ; $f_{i.} = 1/4$ for all i , $f_{.j} = 1/3$ for all j so that $f_{ij} = f_{i.}f_{.j}$ for all i, j . Hence etc.

3. $(x_i, y_j) = (i, j)$ ($i = 1, 2, \dots, 13$; $j = 1, 2, 3, 4$). $f_{ij} = 1/52$ for all i, j .

4. $(x_i, y_j) = (i = 0, 1, 2, \dots; j = 0, 1, 2, \dots, n)$; $f_{ij} = e^{-\mu} \frac{\mu^i}{i!} \binom{n}{j} p^j (1-p)^{n-j}$

5. $f_x(x) = \sin x$ ($0 < x < \pi/2$), $f_y(y) = \sin y$ ($0 < y < \pi/2$)

6. By (6.3.6), $1 = K \int_{x=0}^1 \int_{y=0}^x xy dx dy = K/8$ so that $K = 8$. $f_x(x) =$

$4x^3$ ($0 < x < 1$), $f_y(y) = 4y(1-y^2)$ ($0 < y < 1$).

7. $3/8, 5/24, 5/8, 3/5$. Draw diagrams.

8. $(3x^2 - 8xy + 6y^2)/(6y^2 - 4y + 1)$ ($0 < x < 1, 0 < y < 1$),
 $(3x^2 - 8xy + 6y^2)/(3x^2 - 4x + 2)$ ($0 < x < 1, 0 < y < 1$)

9. $2x$ ($0 < x < 1$), $2(1-y)$ ($0 < y < 1$); $1/(1-y)$ ($y < x < 1$),
 $1/x$ ($0 < y < x$); $1/2$

10. $f_x(x|y) = |y|e^{-y^2x^2}/\sqrt{\pi}$, $f_y(y) = \lambda e^{-\lambda^2y^2}/\sqrt{\pi}$. By (6.5.8) $f(x, y) = \lambda|y|e^{-y^2(x^2+\lambda^2)}/\pi$ so that $f_x(x) = \lambda\pi^{-1} \int_{-\infty}^{\infty} |y|e^{-y^2(x^2+\lambda^2)} dy = 2\lambda\pi^{-1} \times \int_0^{\infty} y e^{-y^2(x^2+\lambda^2)} dy = \lambda/\pi(x^2 + \lambda^2)$.

11. Let X, Y denote the random points. (X, Y) is uniformly distributed over the unit square $R: 0 < x < 1, 0 < y < 1$. The required event is $|X - Y| < k$ or $Y > X - k$ if $X > Y$ and $Y < X + k$ if $Y > X$, i.e. (X, Y) lies in the part R' of R lying between $y = x \pm k$. Hence $R' = 1 - (1 - k)^2 = k(2 - k)$ so that by (6.4.2) the required probability $= R'/R = k(2 - k)$.

12. (X, Y) is uniformly distributed over the unit square $R: 0 < x < 1, 0 < y < 1$ and the event in question is $XY < k$, i.e. (X, Y) lies in $R': xy < k$. Hence $R' = k + k \int_k^1 dx/x = k(1 - \log k)$ etc.

13. (p, q) is uniformly distributed over $R: -1 < p < 1, -1 < q < 1$, and the required event is $p^2 \geq q$, i.e. (p, q) lies in $R': p^2 \geq q$. Hence $R = 4$, $R' = 2 + 2 \int_0^1 p^2 dp = 8/3$, and the answer $= 2/3$.

14. For $0 < x < 1$, $f_x(x) = \int_{-(1-x)}^{1-x} \frac{1}{2} dy = 1 - x$, and for $-1 < x < 0$, $f_x(x) = \int_{-1-x}^{1+x} \frac{1}{2} dy = 1 + x$. Similarly for $0 < y < 1$, $f_y(y) = 1 - y$ and for $-1 < y < 0$, $f_y(y) = 1 + y$.

15. $2\sqrt{a^2 - x^2}/\pi a^2$ ($-a < x < a$), $2\sqrt{a^2 - y^2}/\pi a^2$ ($-a < y < a$), $1/2\sqrt{a^2 - x^2}$ ($-\sqrt{a^2 - x^2} < y < \sqrt{a^2 - x^2}$)

16. Choosing A as the origin, let X, Y be the co-ordinates of P, Q respectively. X, Y are independent, (X, Y) uniformly distributed over the rectangle $R: 0 < x < a, a < y < a + b$. Required event: $X < (a + b)/2, Y > (a + b)/2$ and $Y < X + (a + b)/2$, i.e. (X, Y) lies in the

triangular region $R' : x < (a+b)/2, y > (a+b)/2, y < x + (a+b)/2$.
 $R = ab, R' = \frac{1}{2}b^2$. Required probability $= b/2a$.

17. Let $ABCD$ be a tile which is a parallelogram and diagonal $AC = l$. Take A' on AC such that $AA' = c$, and so $A'C = l - c$. Draw $A'B' \parallel AB, A'D' \parallel AD$. The end of the stick towards C is uniformly distributed over $R : ABCD$, and required event means that the same end lies in $R' : A'B'CD'$. Since $R/R' = (l-c)^2/l^2$ the first result follows.

Draw parallelogram $A_1B_1C_1D_1$ inside the parallelogram $ABCD$ whose sides are parallel to the sides of $ABCD$ and such that the annular region has width $d/2$. The distribution of the centre of the circle is uniform over $R : ABCD$ and the required event means that it lies in $R' : A_1B_1C_1D_1$ so that the answer $= R'/R = (1-d/a)(1-d/b)$.

18. Let X, Y —angles made by OP, OQ respectively with OA, O being the centre of the circle. Then X, Y are independent, and (X, Y) has uniform distribution over the square: $0 < x < 2\pi, 0 < y < 2\pi$. Consider the case $X < Y$. The required event means either $Y < \pi$ or $Y - X > \pi$, i.e. (X, Y) lies in the triangular region: $x > 0, y < \pi, x < y$ or the triangular region: $x > 0, y < 2\pi, y > x + \pi$, each having area $\pi^2/2$. Answer $= 1/2$.

19. Probability mass between the ellipses λ and $\lambda + d\lambda$ is maximum (for fixed $d\lambda$) when $\lambda e^{-\lambda^2/2(1-\rho^2)}$ is maximum, which is so when $\lambda^2 = 1 - \rho^2$. By Sec. 6.4(b) answer $= 1/\sqrt{e}$.

20. Set $u = (x - m_x) \cos \alpha + (y - m_y) \sin \alpha, v = -(x - m_x) \sin \alpha + (y - m_y) \cos \alpha$; $\frac{\partial(u, v)}{\partial(x, y)} = 1$. Define σ_u, σ_v by $\sigma_u^2 = \sigma_x^2 \cos^2 \alpha + \sigma_y^2 \sin^2 \alpha + 2\rho\sigma_x\sigma_y \sin \alpha \cos \alpha, \sigma_v^2 = \sigma_x^2 \sin^2 \alpha + \sigma_y^2 \cos^2 \alpha - 2\rho\sigma_x\sigma_y \sin \alpha \cos \alpha$. Assume $\tan 2\alpha = 2\rho\sigma_x\sigma_y/(\sigma_x^2 - \sigma_y^2)$. Then $\sigma_u^2 + \sigma_v^2 = \sigma_x^2 + \sigma_y^2, \sigma_u^2 - \sigma_v^2 = (\sigma_x^2 - \sigma_y^2) \sec 2\alpha$ so that $4\sigma_u^2\sigma_v^2 = (\sigma_u^2 + \sigma_v^2)^2 - (\sigma_u^2 - \sigma_v^2)^2 = 4\sigma_x^2\sigma_y^2 - (\sigma_x^2 - \sigma_y^2)^2 \tan^2 2\alpha = 4(1 - \rho^2)\sigma_x^2\sigma_y^2$ giving $\sigma_u\sigma_v = \sigma_x\sigma_y \sqrt{1 - \rho^2}$.

Now

$$\frac{(x - m_x)^2}{\sigma_x^2} - 2\rho \frac{(x - m_x)(y - m_y)}{\sigma_x\sigma_y} + \frac{(y - m_y)^2}{\sigma_y^2} = (1 - \rho^2) \left(\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2} \right)$$

By (6.6.1) $f_{u,v}(u, v) = \frac{1}{\sqrt{2\pi}\sigma_u} e^{-u^2/2\sigma_u^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_v} e^{-v^2/2\sigma_v^2}$ etc.

21. $U = X + Y$, $V = X$; $u = x + y$, $v = x$; $x = v$, $y = u - v$; As x , y range from 0 to 1, u ranges from 0 to 2 and v from 0 to 1. By (6.6.2) $f_u(u) = \int_{-\infty}^{\infty} f(x, y) dv = \int_{-\infty}^{\infty} f(v, u - v) dv$. Now $f(v, u - v) = u$ for $0 < v < 1$, $0 < u - v < 1$ and 0 otherwise. Hence if $0 < u < 1$, $f_u(u) = u \int_0^u dv = u^2$, and if $1 < u < 2$, $f_u(u) = u \int_{u-1}^1 dv = u(2 - u)$.

22. u ($0 < u < 1$), $2 - u$ ($1 < u < 2$); $|1 - u|$ ($-1 < u < 1$); $-\log u$ ($0 < u < 1$)

23. $X = U \cos V$, $Y = U \sin V$; $x = u \cos v$, $y = u \sin v$; $\frac{\partial(x, y)}{\partial(u, v)} = u$.

Since X , Y are independent standard normal variates, by (6.6.1) $f_{u, v}(u, v) = u e^{-(x^2 + y^2)/2} / 2\pi = u e^{-u^2/2} / 2\pi$. Hence $f_u(u) = u e^{-u^2/2}$ ($0 < u < \infty$), $f_v(v) = 1/2\pi$ ($0 < v < 2\pi$).

24. Set $X_1 = U \cos V$, $X_2 = U \sin V$; $x_1 = u \cos v$, $x_2 = u \sin v$. As x_1 , x_2 range from 0 to ∞ , u ranges from 0 to ∞ and v from 0 to $\pi/2$. Answer: $2u^3 e^{-u^2}$ ($0 < u < \infty$).

25. Set $U = X_1/X_2$, $V = X_1$; $u = x_1/x_2$, $v = x_1$; u, v both range from 0 to ∞ . $f_u(u) = (1 + u)^{-2}$ ($0 < u < \infty$). $Z = X_2/(X_1 + X_2) = 1/(1 + U)$ has spectrum $(0, 1)$, and $f_z(z) = f_u(u) \left| \frac{du}{dz} \right| = 1$ ($0 < z < 1$).

26. Set $U = XY$, $V = X$; $u = xy$, $v = x$; $x = v$, $y = u/v$. u ranges from $-\infty$ to ∞ and v from -1 to 1 . Let $u > 0$. Then $v = x > 0$ so that v ranges from 0 to 1 only and $\frac{\partial(u, v)}{\partial(x, y)} = -v < 0$ everywhere. Hence

$$f_u(u) = \frac{2u}{\pi} \int_0^1 \frac{e^{-u^2/v^2}}{v^2 \sqrt{1-v^2}} dv = e^{-u^2} / \sqrt{\pi}$$

For $u < 0$, $f_u(u)$ has the same expression.

27. $U = X + Y$ where X , Y are independent Poisson variates having parameters μ_1 , μ_2 respectively. Spectrum of U is given by $u_k = k$ ($k = 0, 1, 2, \dots$), and

$$\begin{aligned} f_{uk} &= \sum_{i+j=k} e^{-(\mu_1+\mu_2)} \frac{\mu_1^i \mu_2^j}{i! j!} = \frac{e^{-(\mu_1+\mu_2)}}{k!} \sum_{i=0}^k \binom{k}{i} \mu_1^i \mu_2^{k-i} \\ &= e^{-(\mu_1+\mu_2)} \frac{(\mu_1 + \mu_2)^k}{k!} \end{aligned}$$

28. $U = X + Y$. Then by (6.6.3)

$$f_u(u) = \frac{\lambda_1 \lambda_2}{\pi^2} \int_{-\infty}^{\infty} \frac{dv}{[(v - \mu_1)^2 + \lambda_1^2][(u - v - \mu_2)^2 + \lambda_2^2]} \quad (-\infty < u < \infty)$$

$$= \frac{\lambda_1 \lambda_2}{\pi^2} \int_{-\infty}^{\infty} \frac{dt}{(t^2 + \lambda_1^2)[(t - a)^2 + \lambda_2^2]} \quad [a = u - \mu_1 - \mu_2]$$

Set

$$\frac{1}{(t^2 + \lambda_1^2)[(t - a)^2 + \lambda_2^2]} = \frac{At + B}{t^2 + \lambda_1^2} - \frac{A(t - a) + C}{(t - a)^2 + \lambda_2^2}$$

Then $aA - B + C = 0$, $A(a^2 - \lambda_1^2 + \lambda_2^2) - 2aB = 0$, $a\lambda_1^2 A + B(a^2 + \lambda_2^2) - \lambda_1^2 C = 1$ giving $A(a^2 + \lambda_1^2 - \lambda_2^2) + 2aC = 0$, $A[a^4 + 2a^2(\lambda_1^2 + \lambda_2^2) + (\lambda_1^2 - \lambda_2^2)^2] = 2a$ or $A[a^2 + (\lambda_1 + \lambda_2)^2][a^2 + (\lambda_1 - \lambda_2)^2] = 2a$. Hence

$$f_u(u) = \frac{A\lambda_1\lambda_2}{2\pi^2} \left[\log \frac{t^2 + \lambda_1^2}{(t - a)^2 + \lambda_2^2} \right]_{-\infty}^{\infty} + \frac{1}{\pi} (\lambda_2 B - \lambda_1 C)$$

First term vanishes, and so $f_u(u) = (\lambda_1 + \lambda_2)/\pi[(u - \mu_1 - \mu_2)^2 + (\lambda_1 + \lambda_2)^2]$.

$X = X_1 + X_2 + \dots + X_n$ is a Cauchy ($n\lambda$, $n\mu$) variate by the first part. Set $\bar{X} = X/n$; $\bar{x} = x/n$. Hence

$$\bar{f}_{\bar{x}}(\bar{x}) = f_x(x) \left| \frac{dx}{d\bar{x}} \right| = \frac{n^2 \lambda}{\pi[(n\bar{x} - n\mu)^2 + n^2 \lambda^2]} = \frac{\lambda}{\pi[(\bar{x} - \mu)^2 + \lambda^2]}$$

29. Set $U = X + Y + Z$, $V = X + Y$, $W = X$; $u = x + y + z$, $v = x + y$, $w = x$; $x = w$, $y = v - w$, $z = u - v$; $\frac{\partial(u, v, w)}{\partial(x, y, z)} = -1$. Hence

$$f_u(u) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(w, v - w, u - v) dv dw$$

$$= \frac{6}{(1 + u)^4} \int_0^u dv \int_0^v dw = \frac{3u^2}{(1 + u)^4} \quad (0 < u < \infty)$$

30. Set $Y_1 = X_1$, $Y_2 = X_1 X_2$, ..., $Y_n = X_1 X_2 \dots X_n$; $y_1 = x_1$, $y_2 = x_1 x_2$, ..., $y_n = x_1 x_2 \dots x_n$ so that $x_1 = y_1$, $x_2 = y_2/y_1$, ..., $x_n = y_n/y_{n-1}$. Density function of Y_n is

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y_1, y_2/y_1, \dots, y_n/y_{n-1}) \frac{dy_1 dy_2 \dots dy_{n-1}}{y_1 y_2 \dots y_{n-1}}$$

$$= \iint \dots \int \frac{dy_1 dy_2 \dots dy_{n-1}}{y_1 y_2 \dots y_{n-1}}$$

the range of integration being $y_2 < y_1 < 1$, $y_3 < y_2 < 1$, ..., $y_n < y_{n-1} < 1$.

31. For $a < x < b$, $(X > x) = (X_1 > x, X_2 > x, \dots, X_n > x)$ so that $P(X > x) = P(X_1 > x) \dots P(X_n > x) = (b-x)^n / (b-a)^n$ or $F(x) = 1 - (b-x)^n / (b-a)^n$. Hence $f(x) = F'(x) = n(b-x)^{n-1} / (b-a)^n$.

7.14

1. X —number of white balls. (a) X is binomial $(n, N_1 / (N_1 + N_2))$ so that $E(X) = nN_1 / (N_1 + N_2)$, (b) $P(X=i) = \binom{N_1}{i} \binom{N_2}{n-i} / \binom{N_1+N_2}{n}$.

Then $\sum \binom{N_1}{i} \binom{N_2}{n-i} / \binom{N_1+N_2}{n} = 1$, the summation being over all possible values of i . $E(X) = \sum i \binom{N_1}{i} \binom{N_2}{n-i} / \binom{N_1+N_2}{n} = \frac{nN_1}{N_1+N_2} \times \sum \binom{N_1-1}{i-1} \binom{N_2}{n-i} / \binom{N_1+N_2-1}{n-1} = \frac{nN_1}{N_1+N_2}$.

2. $2a^2/3, a$

3. $2/\pi; 2/\pi \sqrt{1-y^2} \ (0 < y < 1)$

4. See Ex. 7 Sec. 5.10. $\sum_{i=0}^n i \binom{2n-i}{n} \left(\frac{1}{2}\right)^{2n-i} = \sum_{i=0}^n \{n - (n-i)\} \times \binom{2n-i}{n} \left(\frac{1}{2}\right)^{2n-i} = n - (n+1) \sum_{i=0}^{n-1} \binom{2n-i}{n+1} \left(\frac{1}{2}\right)^{2n-i}$. Replacing n by $n+1$ in the identity $\sum_{i=0}^n \binom{2n-i}{n} \left(\frac{1}{2}\right)^{2n-i} = 1$ we get $1 = \sum_{i=0}^{n+1} \binom{2n-i+2}{n+1} \times \left(\frac{1}{2}\right)^{2n-i+2} = \binom{2n+2}{n+1} \left(\frac{1}{2}\right)^{2n+2} + \binom{2n+1}{n+1} \left(\frac{1}{2}\right)^{2n+1} + \sum_{i=0}^{n-1} \binom{2n-i}{n+1} \left(\frac{1}{2}\right)^{2n-i} = \binom{2n+2}{n+1} \left(\frac{1}{2}\right)^{2n+1} + \sum_{i=0}^{n-1} \binom{2n-i}{n+1} \left(\frac{1}{2}\right)^{2n-i}$. Answer $= (2n+1) \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} - 1$.

5. See Ex. 8 Sec. 5.10 and Ex. 8 below.

6. $\Gamma(l + \frac{1}{2}) / \Gamma(l)$

7. $(a+b)/2, (b-a)^2/12$

8. Differentiate the identity $\sum_{i=0}^{\infty} x^i = (1-x)^{-1}$ once and twice and

put $x = \mu/(1 + \mu)$. Using these results

$$m = \frac{\mu}{(1 + \mu)^2} \sum i \left(\frac{\mu}{1 + \mu} \right)^{i-1} = \mu$$

$$E\{X(X-1)\} = \frac{\mu^2}{(1 + \mu)^3} \sum i(i-1) \left(\frac{\mu}{1 + \mu} \right)^{i-2} = 2\mu^2$$

so that $\sigma^2 = 2\mu^2 - \mu(\mu-1) = \mu(\mu+1)$.

9. $e^2 - e$

10. 3, 15, -86

11. 1, 1/5, 0

16. $\frac{l(l+1) \cdots (l+k-1)}{(l+m)(l+m+1) \cdots (l+m+k-1)}, \frac{lm}{(l+m)^2(l+m+1)}$

17. $\sqrt{e}, 1/e, \sqrt{e(e-1)}, (1 - e^{-3/2})/\sqrt{e-1}$

18. $f_{i+1}/f_i - 1 = (\mu + 1 - i)/(i + 1)$. Hence argue.

19. $1/(1 - t/a), k!/a^k$

20. $\mu_{2k+1} = 0, \mu_{2k} = a^{2k}/(2k+1)$

21. $\alpha_1 = \mu, \alpha_2 = \mu + 2\mu^2, \alpha_3 = \mu + 6\mu^2 + 6\mu^3, \alpha_4 = \mu + 14\mu^2 + 36\mu^3 + 24\mu^4; \kappa_1 = \mu, \kappa_2 = \mu + \mu^2, \kappa_3 = \mu + 3\mu^2 + 2\mu^3, \kappa_4 = \mu + 7\mu^2 + 12\mu^3 + 6\mu^4; \gamma_1 = (2\mu + 1)/\sqrt{\mu(\mu + 1)}, \gamma_2 = (6\mu^2 + 6\mu + 1)/\mu(\mu + 1)$

22. $\chi(t) = e^{i\mu t}/(1 + \lambda^2 t^2); \kappa_1 = \mu, \kappa_2 = 2\lambda^2, \kappa_3 = 0, \kappa_4 = 12\lambda^4; m = \mu, \sigma = \sqrt{2\lambda}, \gamma_1 = 0, \gamma_2 = 3$

23. 1, 1, 1

24. 2

25. 4, 10, 3

26. $\lambda, \lambda \log_e 2$

8.16

1. $2a^n/(n+1)(n+2)$

$$\begin{aligned} 2. E\{\max(X, Y)\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max(x, y) f(x, y) dx dy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} 2 \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} x e^{-\frac{(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)}} dx \end{aligned}$$

$$3. E\{\min(|X_1|, \dots, |X_n|)\}$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \min(|x_1|, \dots, |x_n|) f(x_1, \dots, x_n) dx_1 \dots dx_n \\ &= n 2^{n-1} \int_{-\infty}^{\infty} \int_{|x_1|}^{\infty} \dots \int_{|x_1|}^{\infty} |x_1| \phi(x_1) \dots \phi(x_n) dx_1 \dots dx_n \quad [\phi'(x) = \phi(x)] \\ &= n 2^n \int_0^{\infty} \int_{x_1}^{\infty} \dots \int_{x_1}^{\infty} x_1 \phi(x_1) \dots \phi(x_n) dx_1 \dots dx_n \\ &= n 2^n \int_0^{\infty} x_1 \phi(x_1) [1 - \phi(x_1)]^{n-1} dx_1 \\ &= 2^n \int_0^{\infty} [1 - \phi(x_1)]^n dx_1 \end{aligned}$$

4. Spectrum : $(x_i, y_j) = (i, j)$ ($i, j = 0, 1, 2$) ; $f_{ij} = 1/6$ for all i, j except that $f_{12} = f_{21} = f_{22} = 0$. $m_x = m_y = 2/3$, $\alpha_{x2} = \alpha_{y2} = 1$, $\alpha_{11} = 1/6$. Hence $\sigma_x = \sigma_y = \sqrt{5/3}$, $\rho = -1/2$. Regression lines : $x + 2y = 2$, $2x + y = 2$.

5. $y - \frac{7}{9} = \frac{17}{60}(x - 4)$, $x - 4 = \frac{153}{50}(y - \frac{7}{9})$; $\rho = 17\sqrt{30}/100$; $|\rho|$ is a measure of goodness of fit.

6. X —number of heads, Y —longest run of heads. Write down all the 16 event points and the corresponding values of X, Y . Spectrum : $(x_i, y_j) = (i, j)$ ($i, j = 0, 1, 2, 3, 4$) ; $f_{00} = 1/16$, $f_{11} = 4/16$, $f_{21} = 3/16$, $f_{22} = 3/16$, $f_{32} = 2/16$, $f_{33} = 2/16$, $f_{44} = 1/16$, $f_{ij} = 0$ for other values of i, j . $m_x = 2$, $m_y = 27/16$, $\sigma_x^2 = 1$, $\sigma_y^2 = 247/256$, $\rho = 14/\sqrt{247}$.

$$7. \sqrt{27/73}$$

$$8. m_x = 3/2, m_y = 1, \alpha_{11} = 3/2, \mu_{11} = 0$$

$$9. Y = X - 1$$

10. $7/12, 7/12, \sqrt{11}/12, \sqrt{11}/12, -1/11$. Regression lines :

$$y - \frac{7}{12} = -\frac{1}{11}\left(x - \frac{7}{12}\right), y - \frac{7}{12} = -11\left(x - \frac{7}{12}\right)$$

Regression curves :

$$y = \frac{3x+2}{3(2x+1)}, x = \frac{3y+2}{3(2y+1)}$$

11. Regression curves :

$$y = \frac{9x^2 - 16x + 9}{6(3x^2 - 4x + 2)}, \quad x = \frac{36y^2 - 32y + 9}{12(6y^2 - 4y + 1)}$$

Regression lines :

$$y - \frac{2}{3} = -\frac{30}{67} \left(x - \frac{5}{12} \right), \quad x - \frac{5}{12} = -\frac{15}{32} \left(y - \frac{2}{3} \right)$$

13. The point of intersection is $(3, \frac{1}{2})$ which gives $m_x = 3, m_y = \frac{1}{2}$, $\rho\sigma_y/\sigma_x = -1/6, \rho\sigma_x/\sigma_y = -2/3$ so that $\rho^2 = 1/9$ or $\rho = -1/3$ as $\rho < 0$.

14. $n = 13$. X_i - point in the i th drawing ($i = 1, 2, \dots, n$). From symmetry X_i takes values 1, 2, 3, 4 each with probability $1/4$ so that $E(X_i) = 5/2$. $S_n = X_1 + \dots + X_n$ is the total number of points; $E(S_n) = E(X_1) + \dots + E(X_n) = n \cdot \frac{5}{2} = 65/2$.

15. $a^2/8$

16. $0 \leq E[(X - kY)^2] = E(X^2) - 2kE(XY) + k^2E(Y^2)$. Putting $k = E(XY)/E(Y^2)$ the inequality follows, provided $E(Y^2) \neq 0$. If $E(Y^2) = 0$, then Y has a one-point distribution at $y = 0$ so that $E(XY) = 0$ and the equality holds.

The inequality for X^*, Y^* gives the second conclusion.

17. $U = aX + bY, V = cY$; $m_u = am_x + bm_y, m_v = cm_y$; $\sigma_u^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_x\sigma_y\rho, \sigma_v = c\sigma_y$ since $c > 0$. $(U - m_u)(V - m_v) = ac \times (X - m_x)(Y - m_y) + bc(Y - m_y)^2$ so that $\text{cov}(U, V) = ac\sigma_x\sigma_y\rho + bc\sigma_y^2$ etc.

$$18. \frac{a_1 a_2 \sigma_x^2 + b_1 b_2 \sigma_y^2}{\sqrt{(a_1^2 \sigma_x^2 + b_1^2 \sigma_y^2)(a_2^2 \sigma_x^2 + b_2^2 \sigma_y^2)}}$$

21. $X = U \cos \alpha - V \sin \alpha, Y = U \sin \alpha + V \cos \alpha$; Since U, V are uncorrelated $\sigma_x^2 = \sigma_u^2 \cos^2 \alpha + \sigma_v^2 \sin^2 \alpha, \sigma_y^2 = \sigma_u^2 \sin^2 \alpha + \sigma_v^2 \cos^2 \alpha$ and $\tan 2\alpha = 2\rho\sigma_x\sigma_y/(\sigma_x^2 - \sigma_y^2)$. Hence $\sigma_x^2 + \sigma_y^2 = \sigma_u^2 + \sigma_v^2, \sigma_x^2 - \sigma_y^2 = (\sigma_u^2 - \sigma_v^2) \cos 2\alpha$. Then $(\sigma_u^2 - \sigma_v^2)^2 = (\sigma_x^2 - \sigma_y^2)^2 \sec^2 2\alpha = (\sigma_x^2 - \sigma_y^2)^2 + 4\rho^2 \sigma_x^2 \sigma_y^2$ or $(\sigma_u^2 + \sigma_v^2)^2 - 4\sigma_u^2 \sigma_v^2 = (\sigma_x^2 + \sigma_y^2)^2 - 4(1 - \rho^2)\sigma_x^2 \sigma_y^2$ or $\sigma_u \sigma_v = \sigma_x \sigma_y \sqrt{1 - \rho^2}$.

22. $\rho = \pm 1$. $\sigma_v^2 = \sin^2 \alpha \sigma_x^2 + \cos^2 \alpha \sigma_y^2 - 2\rho \sin \alpha \cos \alpha \sigma_x \sigma_y = (\sin \alpha \sigma_x - \rho \cos \alpha \sigma_y)^2 = 0$ if $\tan \alpha = \rho \sigma_y / \sigma_x$.

23. Show that the covariance vanishes.

24. Define random variables X_1, X_2, \dots, X_n on the event space of n trials by: $X_i = 0$ or 1 corresponding to 'failure' or 'success' in the i th trial ($i = 1, 2, \dots, n$). X_i is binomial $(1, p_i)$ so that $m_i = p_i, \sigma_i^2 = p_i q_i$.

$S_n = X_1 + \cdots + X_n$ denotes the number of successes in n trials, and the trials being independent, X_1, \dots, X_n are mutually independent. By (8.5.2) and (8.5.3) the results follow.

25. There are $N_1 = 16$ aces, kings, queens and jacks and $N_2 = 36$ other cards in the pack from which $n = 13$ are drawn without replacement. Hence by the example of Sec. 8.5, $M_n = 4$, $\Sigma_n^2 = 36/17$.

26. Define random variables X_1, X_2, \dots, X_n on the event space of n drawings by: X_i - number on the i th ticket ($i = 1, 2, \dots, n$) so that $S_n = X_1 + \cdots + X_n$ is the sum of numbers on the tickets drawn. From symmetry X_1, \dots, X_n all have the same distribution. Now X_1 takes values $1, 2, \dots, N$ each with probability $1/N$ so that $m_1 = E(X_1) = (N+1)/2$, $\sigma_1^2 = \text{var}(X_1) = (N^2 - 1)/12$. Hence $m_i = (N+1)/2$, $\sigma_i^2 = (N^2 - 1)/12$ ($i = 1, 2, \dots, n$). By (8.5.2) $M_n = n(N+1)/2$. Consider now the distribution of (X_1, X_2) . Spectrum: (i, j) ($i, j = 1, 2, \dots, N$) and $P(X_1 = i, X_2 = j) = 1/N(N-1)$ if $i \neq j$ and 0 if $i = j$. Hence

$$E(X_1 X_2) = \frac{1}{N(N-1)} \sum_{j=1}^N \sum_{i=1}^N ij - \sum_{i=1}^N i^2 = (N+1)(3N+2)/12$$

Then $\text{cov}(X_1, X_2) = -(N+1)/12$. From symmetry $\text{cov}(X_i, X_j) = -(N+1)/12$ ($i \neq j$). Then by (8.5.3)

$$\Sigma_n^2 = \frac{n(N^2 - 1)}{12} \left(1 - \frac{n-1}{N-1} \right)$$

27. X - number of matches. By (3.2.6) we have the identity

$$\sum_{i=0}^n \sum_{k=0}^{n-i} \frac{(-1)^k}{i! k!} = 1$$

$$E(X) = \sum_{i=1}^n \sum_{k=0}^{n-i} \frac{(-1)^k}{(i-1)! k!} = \sum_{i=0}^{n-1} \sum_{k=0}^{n-1-i} \frac{(-1)^k}{i! k!} = 1$$

$$E\{X(X-1)\} = \sum_{i=2}^n \sum_{k=0}^{n-i} \frac{(-1)^k}{(i-2)! k!} = \sum_{i=0}^{n-2} \sum_{k=0}^{n-2-i} \frac{(-1)^k}{i! k!} = 1$$

so that $\text{var}(X) = 1$.

Another method. On the event space of n drawings we define random variables X_1, X_2, \dots, X_n by: $X_i = 1$ or 0 corresponding to match or no match in the i th drawing, so that $S_n = X_1 + \cdots + X_n$ is

the number of matches in n drawings. From symmetry X_1, \dots, X_n are identically distributed. $P(X_1 = 1) = 1/n$, and hence $m_1 = E(X_1) = 1/n$, $\sigma_1^2 = \text{var}(X_1) = (n-1)/n^2$. Hence $m_i = 1/n$, $\sigma_i^2 = (n-1)/n^2$ ($i = 1, 2, \dots, n$). By (8.5.2) $M_n = 1$. The two-dimensional variate (X_1, X_2) takes values $(0, 0)$, $(1, 0)$, $(0, 1)$, $(1, 1)$ and $P(X_1 = 1, X_2 = 1) = 1/n(n-1)$, so that $E(X_1 X_2) = 1/n(n-1)$ giving $\text{cov}(X_1, X_2) = 1/n^2(n-1)$. From symmetry $\text{cov}(X_i, X_j) = 1/n^2(n-1)$ ($i \neq j$). By (8.5.3) $\Sigma_n^2 = 1$.

28. Let $\chi(t)$ be the characteristic function of each of the random variables X_1, \dots, X_n and $K(t)$ that of their sum. By (8.8.3) $\{\chi(t)\}^n = K(t) = e^{imt - \sigma^2 t^2 / 2}$ or $\chi(t) = e^{imt/n - \sigma^2 t^2 / 2n}$ which corresponds to a normal $(m/n, \sigma/\sqrt{n})$ distribution.

30. See random walk problem, Sec. 8.9. Here $r = 5$, $n = 9$, $p = q = \frac{1}{2}$, $x'_i = 2i - 4 = 0$ if $i = 2$ so that by (8.9.1) the required probability $= 9/128$.

31. $\chi(t, u) = p_{00} + p_{10}e^{it} + p_{01}e^{iu} + p_{11}e^{i(t+u)}$. If $p_{00}p_{11} = p_{01}p_{10}$, $\chi(t, u) = p_{11}^{-1} [p_{00}p_{10} + p_{11}p_{10}e^{it} + p_{11}p_{01}e^{iu} + p_{11}^2 e^{i(t+u)}] = p_{11}^{-1} (p_{01} + p_{11}e^{it})(p_{10} + p_{11}e^{iu})$; $\chi_x(t) = \chi(t, 0) = p_{11}^{-1} (p_{10} + p_{11}) \times (p_{01} + p_{11}e^{it})$, $\chi_y(u) = \chi(0, u) = p_{11}^{-1} (p_{01} + p_{11})(p_{10} + p_{11}e^{iu})$. $\chi(t, u) = \chi_x(t) \chi_y(u)$ if $(p_{10} + p_{11})(p_{01} + p_{11}) = p_{11}$ or $p_{00}p_{11}^2 + p_{11}(p_{10} + p_{01}) + p_{11}^2 = p_{11}$ or if $p_{00} + p_{10} + p_{01} + p_{11} = 1$ which is true.

32. By (8.13.17), (8.13.18), (8.13.19) $\sigma^2(V_y) = E(V_y^2) = E\{(Y - c_0^* - c_1^* X)V_y\} = E(YV_y) = E\{Y(Y - c_0^* - c_1^* X)\} = a_{02} - c_0^* a_{01} - c_1^* a_{11}$.

33. See Ex. 4 above. $a_{x3} = 5/3$, $a_{x4} = 3$, $a_{21} = 1/6$. By (8.14.4) the normal equations are $c_0^* + \frac{2}{3}c_1^* + c_2^* = \frac{2}{3}$, $\frac{2}{3}c_0^* + c_1^* + \frac{5}{3}c_2^* = \frac{1}{6}$, $c_0^* + \frac{5}{3}c_1^* + 3c_2^* = \frac{1}{6}$ which give $c_0^* = 1$, $c_1^* = -\frac{1}{2}$, $c_2^* = 0$. Hence the required parabola reduces to the straight line $y = 1 - \frac{1}{2}x$. Since this is also the regression line of Y on X , a measure goodness of fit is $|\rho| = 1/2$.

34. X is normal $(0, \sigma_x)$. $a_{x2} = \sigma_x^2$, $a_{x3} = 0$, $a_{x4} = 3\sigma_x^4$, $a_{11} = \rho\sigma_x\sigma_y$.

$$a_{21} = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 y e^{-Q/2(1-\rho^2)} dx dy$$

where $Q = \frac{x^2}{\sigma_x^2} - 2\rho \frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}$ so that

$$a_{21} = \frac{\sigma_x^2 \sigma_y}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 y e^{-(x^2 - 2\rho xy + y^2)/2(1-\rho^2)} dx dy$$

$$\begin{aligned}
&= \frac{\sigma_x^2 \sigma_y}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\
&\quad \times \left\{ \frac{1}{\sqrt{2\pi} \sqrt{1-\rho^2}} \int_{-\infty}^{\infty} y e^{-(y-\rho x)^2/2(1-\rho^2)} dy \right\} \\
&= \frac{\rho \sigma_x^2 \sigma_y}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 e^{-x^2/2} dx = 0
\end{aligned}$$

Normal equations: $c_0^* + c_2^* \sigma_x^2 = 0$, $c_1^* \sigma_x^2 = \rho \sigma_x \sigma_y$, $c_0^* \sigma_x^2 + 3c_2^* \sigma_x^4 = 0$ which give $c_0^* = 0$, $c_1^* = \rho \sigma_y / \sigma_x$, $c_2^* = 0$. Hence the regression parabola reduces to the straight line $y = \rho \sigma_y x / \sigma_x$. Here the regressions for the means are linear.

35. $f_x(x) = \frac{1}{2} x^2 e^{-x}$, $f_y(y) = 3/(y+1)^4$, $\sigma_y^2 = 3/4$, $m_y(x) = 1/x$. σ_{yx}^2
 $= \frac{1}{2} \int_0^{\infty} \int_0^{\infty} \left(y - \frac{1}{x}\right)^2 x^3 e^{-x(y+1)} dx dy = \frac{1}{2}$. By (8.15.6) $\eta_{yx} = 1/\sqrt{3}$.

36. Write $E\{g(X)\} = a$. By (8.15.7) $E\{Y - m_y(X)\} = 0$. Then $\text{cov}\{g(X), Y - m_y(X)\}$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{g(x) - a\} \{y - m_y(x)\} f(x, y) dx dy \\
&= \int_{-\infty}^{\infty} \{g(x) - a\} f_x(x) dx \int_{-\infty}^{\infty} \{y - m_y(x)\} f_y(y|x) dy = 0
\end{aligned}$$

since the latter integral vanishes. $Y - c_0^* - c_1^* X = \{Y - m_y(X)\} + \{m_y(X) - c_0^* - c_1^* X\}$. By (8.13.15) and (8.15.7), the means of $Y - m_y(X)$ and $m_y(X) - c_0^* - c_1^* X$ are both zero so that $E\{(Y - c_0^* - c_1^* X)^2\} = E\{(Y - m_y(X))^2\} + E\{(m_y(X) - c_0^* - c_1^* X)^2\} + 2 \text{cov}\{Y - m_y(X), m_y(X) - c_0^* - c_1^* X\}$. By the first part the last term is zero, and the second result follows from (8.13.9) and (8.15.6). When $\eta_{yx} = |\rho|$, $m_y(x) = c_0^* + c_1^* x$, i.e. the regression for the mean of Y is linear.

9.4

$$1. f(x) = \frac{e^{-x/2\sigma^2} x^{n/2-1}}{2^{n/2} \sigma^n \Gamma(n/2)} \quad (0 < x < \infty)$$

$$2. f(u) = \sqrt{2\alpha} m^{-3/2} u^{1/2} e^{-\alpha\beta u/m} \quad (0 < u < \infty)$$

3. See Ex. 1 Sec. 6.6.

$$4. \gamma_1 = \sqrt{8/n}, \gamma_2 = 12/n$$

5. $Z = X + Y$. Since X, Y are independent. $\chi_z(t) = \chi_x(t)\chi_y(t)$. Given $\chi_x(t) = (1 - 2it)^{-m/2}$ $\chi_z(t) = (1 - 2it)^{-(m+n)/2}$ it follows that $\chi_y(t) = (1 - 2it)^{-n/2}$ etc.

6. For $n > 1$ the mean is zero, and hence

$$\begin{aligned} \sigma^2 &= \frac{1}{\sqrt{n} B(\frac{1}{2}, \frac{1}{2}n)} \int_{-\infty}^{\infty} \frac{t^2 dt}{(1 + t^2/n)^{(n+1)/2}} \\ &= \frac{n}{B(\frac{1}{2}, \frac{1}{2}n)} \int_0^{\infty} \frac{x^{\frac{1}{2}} dx}{(1 + x)^{(n+1)/2}} \end{aligned}$$

This integral converges only for $n > 2$, and in that case

$$\sigma^2 = nB(\frac{3}{2}, \frac{1}{2}n - 1)/B(\frac{1}{2}, \frac{1}{2}n) = n/(n - 2)$$

7. For $n > 1$ the mean is zero, and hence

$$\begin{aligned} \mu_4 &= \frac{1}{\sqrt{n} B(\frac{1}{2}, \frac{1}{2}n)} \int_{-\infty}^{\infty} \frac{t^4 dt}{(1 + t^2/n)^{(n+1)/2}} \\ &= \frac{n^2}{B(\frac{1}{2}, \frac{1}{2}n)} \int_0^{\infty} \frac{x^{3/2} dx}{(1 + x)^{(n+1)/2}} \\ &= n^2 B(\frac{5}{2}, \frac{1}{2}n - 2)/B(\frac{1}{2}, \frac{1}{2}n) = 3n^2/(n - 2)(n - 4) \quad [\text{if } n > 4] \end{aligned}$$

By Ex. 4 $\gamma_2 = \mu_4/\sigma^4 - 3 = 6/(n - 4)$.

$$10. \beta_2(\frac{1}{2}m, \frac{1}{2}n)$$

10.3

3. Number of heads is binomial $(2000, \frac{1}{2})$ so that $m = 1000$, $\sigma^2 = 500$. Take $\varepsilon = 100$ in (10.1.1).

$$4. 17/625, 1/4$$

5. Put $\varepsilon = n$ in (10.1.1).
 6. $\Sigma_n^2 = \sigma_1^2 + \dots + \sigma_n^2 \leq nA^2$ so that $\Sigma_n = o(n)$.
 7. Take X_1, X_2, \dots, X_n to be mutually independent, each binomial $(1, p)$.
 8. $p_i q_i \leq \left(\frac{p_i + q_i}{2}\right)^2 = \frac{1}{4}$ so that $\Sigma_n^2 \leq n/4$ and $\Sigma_n = o(n)$.

11.3

1. Number of heads is binomial ($n=2000$, $p=1/2$); $np=1000$, $\sqrt{npq}=10\sqrt{5}$. Taking $a=-2\sqrt{5}$, $b=2\sqrt{5}$ in (11.1.5), the answer $=2\phi(2\sqrt{5})-1 \simeq 0.999992$ (See Fisher and Yates: Statistical Tables).

2. If the die is thrown n times, X_n , the frequency ratio of sixes is binomial ($n, \frac{1}{6}$). Given $P(|X_n/n - \frac{1}{6}| < .01) = .99$ or $P\left(\left|\frac{X_n - n/6}{\sqrt{5n/6}}\right| < .012\sqrt{5n}\right) = .99$, by Table I, $.012\sqrt{5n} = 2.58$ or $n = 9245$.

3. Let us consider the continuous case only, the discrete case being similar. For any $\tau > 0$

$$\begin{aligned} \int_{|x-m_k| \geq \tau \Sigma_n} (x-m_k)^2 f_k(x) dx &\leq \frac{1}{\tau \Sigma_n} \int_{|x-m_k| \geq \tau \Sigma_n} |x-m_k|^3 f_k(x) dx \\ &\leq \frac{1}{\tau \Sigma_n} \int_{-\infty}^{\infty} |x-m_k|^3 f_k(x) dx \\ &= \frac{1}{\tau \Sigma_n} E(|X_k - m_k|^3) \end{aligned}$$

Hence if Liapounoff's condition is satisfied,

$$0 \leq \text{R.H.S. of (11.2.1)}$$

$$\leq \frac{1}{\tau} \lim_{n \rightarrow \infty} \frac{1}{\Sigma_n^3} \sum_{k=1}^n E(|X_k - m_k|^3) = 0$$

so that R.H.S. of (11.2.1) is 0, i.e. Lindeberg's condition is fulfilled.

4. $\Sigma_n^2 = \sigma_1^2 + \dots + \sigma_n^2$, $\Sigma_n > 0$, $\{\Sigma_n\}$ is monotonic increasing. Hence either $\Sigma_n \rightarrow \infty$ or $\{\Sigma_n\}$ is bounded. In the latter case $\Sigma_n \leq A$, a constant for all n . Consider the continuous case. For any $\tau > 0$

$$\int_{|x-m_1| \geq \tau \Sigma_n} (x-m_1)^2 f_1(x) dx \geq \int_{|x-m_1| \geq \tau A} (x-m_1)^2 f_1(x) dx = k \text{ (say)}$$

Choose τ so small that $k \neq 0$ and so $k > 0$, which is always possible. Then for this τ , R.H.S. of (11.2.1) $\geq k/A^2 > 0$ which contradicts Lindeberg's condition.

5. $X - \gamma(n)$ variate. Then $Y = (X - n)/\sqrt{n}$ is the corresponding standardised variate which has spectrum $(-\sqrt{n}, \infty)$. Set $y = (x - n)/\sqrt{n}$ or $x = y\sqrt{n} + n$. $f_y(y) = n^{n+\frac{1}{2}} e^{-n} \{e^{-y/\sqrt{n}} (1 + y/\sqrt{n})\}^n / n! (1 + y/\sqrt{n})$. Now as $n \rightarrow \infty$, $1 + y/\sqrt{n} \rightarrow 1$ and by (3.1.12) $n^{n+\frac{1}{2}} e^{-n} / n! \rightarrow 1/\sqrt{2\pi}$.

$$e^{-y/\sqrt{n}} = 1 - \frac{y}{\sqrt{n}} + \frac{y^2}{2n} - \frac{\xi y^3}{6n^{3/2}}$$

where $\xi = \xi(n, y)$ is such that $0 < \xi \leq A$, constant, remembering that $y > -\sqrt{n}$.

$$\begin{aligned} & \{e^{-y/\sqrt{n}} (1 + y/\sqrt{n})\}^n \\ &= \left\{1 - \frac{y}{\sqrt{n}} + \left(\frac{1}{2} - \frac{\xi}{6}\right) \frac{y^2}{n} - \frac{\xi y^3}{6n^{3/2}}\right\}^n \rightarrow e^{-y^2/2} \text{ as } n \rightarrow \infty \end{aligned}$$

so that $f_y(y) \rightarrow \phi(y)$ as $n \rightarrow \infty$. Then

$$\begin{aligned} \lim_{n \rightarrow \infty} F_y(y) &= \lim_{n \rightarrow \infty} \int_{-\infty}^y f_y(y) dy = \int_{-\infty}^y \lim_{n \rightarrow \infty} f_y(y) dy \\ &= \int_{-\infty}^y \phi(y) dy = \Phi(y) \end{aligned}$$

6. (a) $X_n - \chi^2(n)$ variate. Then c.f. of $(X_n - n)/\sqrt{2n}$, $\chi_n^*(t) = e^{-it\sqrt{n/2}} (1 - it\sqrt{2/n})^{-n/2} = \{e^{it\sqrt{2/n}} (1 - it\sqrt{2/n})\}^{-n/2}$. Now

$$e^{it\sqrt{2/n}} = 1 + it\sqrt{\frac{2}{n}} - \frac{t^2}{n} + \theta \frac{t^3}{6} \left(\frac{2}{n}\right)^{3/2}$$

where θ is a complex quantity such that $|\theta| < 1$. Then

$$\chi_n^*(t) = \left\{1 + \frac{t^2}{n} + \left(\frac{i}{2} + \frac{\theta}{6}\right) t^3 \left(\frac{2}{n}\right)^{3/2} - \frac{2i\theta t^4}{3n^2}\right\}^{-n/2} \rightarrow e^{-t^2/2}$$

(b) If X_n is Poisson- n variate, c.f. of $(X_n - n)/\sqrt{n}$, $\chi_n^*(t) = e^{-it/\sqrt{n}} e^{n(e^{it/\sqrt{n}} - 1)} = e^{n(e^{it/\sqrt{n}} - 1 - it/\sqrt{n})}$. Now

$$e^{it/\sqrt{n}} = 1 + \frac{it}{\sqrt{n}} - \frac{t^2}{2n} + \theta \frac{t^3}{6n^{3/2}}$$

where θ is a complex quantity such that $|\theta| < 1$. Hence

$$n(e^{it/\sqrt{n}} - 1 - it/\sqrt{n}) = -t^2/2 + \theta t^3/6\sqrt{n} \rightarrow -t^2/2 \text{ and } \chi_n^*(t) \rightarrow e^{-t^2/2}$$

12.6

1. 4.86, 4, 4
2. Classes : 15-20, 20-25, ..., 40-45 ; 25.90, 45.44
3. 1.08, 0.92
4. 8.20, 6.72, 0.12, -0.58, 9, 8, 2, 11
5. 0.79197, 0.01175, 0.26, -0.54

13.5

The following results will be useful in the sequel. X_1, X_2, \dots, X_n are mutually independent, each having the distribution of the population.

$$E(X_i - m) = 0$$

$$E\{\Sigma(X_i - m)(X_j - m)\} = \Sigma E\{(X_i - m)(X_j - m)\} = n\sigma^2$$

$$E\{\Sigma(X_i - m)(X_j - m)(X_k - m)\} = \Sigma E\{(X_i - m)(X_j - m)(X_k - m)\} = n\mu_3$$

$$E\{\Sigma(X_i - m)(X_j - m)(X_k - m)(X_l - m)\} = \Sigma E\{(X_i - m)(X_j - m) \times$$

$$(X_k - m)(X_l - m)\} = \Sigma E\{(X_i - m)^4\} + \left(\frac{4}{2}\right) \sum_{i < j} E\{(X_i - m)^2(X_j - m)^2\}$$

$$= n\mu_4 + 3n(n-1)\sigma^4$$

$$E\{\Sigma(X_i - m)^2(X_j - m)\} = \Sigma E\{X_i - m\}^2(X_j - m) = n\mu_3$$

$$E\{\Sigma(X_i - m)^2(X_j - m)(X_k - m)\} = E\{\Sigma(X_i - m)^2(X_j - m)^2\}$$

$$= n\mu_4 + n(n-1)\sigma^4$$

$$E\{\Sigma(X_i - m)^3(X_j - m)\} = n\mu_4$$

1. $E(\bar{X}) = m$, $\sigma(\bar{X}) = \sigma/\sqrt{n}$. $\bar{X} - m = n^{-1} \sum (X_i - m)$ so that $(\bar{X} - m)^3 = n^{-3} \sum (X_i - m)(X_j - m)(X_k - m)$ and $(\bar{X} - m)^4 = n^{-4} \sum (X_i - m)(X_j - m) \times (X_k - m)(X_l - m)$. By the above results, $\mu_3(\bar{X}) = \mu_3/n^3$, $\mu_4(\bar{X}) = \mu_4/n^3 + 3(n-1)\sigma^4/n^3$.

2. $S^2 = n^{-1} \sum (X_i - m)^2 - (\bar{X} - m)^2$ so that $S^4 = n^{-2} \sum (X_i - m)^2 \times (X_j - m)^2 - 2n^{-2} \sum (X_i - m)^2 (X_j - m) + (\bar{X} - m)^4$. By the above relations and using the values of $\mu_4(\bar{X})$ from the working of Ex. 1, $E(S^4) = (n-1)^2 \mu_4/n^3 + (n-1)(n^2 - 2n + 3)\sigma^4/n^3$. By (13.3.2) $\sigma^2(S^2) = E(S^4) - \{E(S^2)\}^2 = (n-1)^2 \mu_4/n^3 - (n-1)(n-3)\sigma^4/n^3$.

3. $M_3 = n^{-1} \sum \{(X_i - m)^3 - 3(X_i - m)(\bar{X} - m) + 3(X_i - m)(\bar{X} - m)^2 - (\bar{X} - m)^3\} = n^{-1} \sum (X_i - m)^3 - 3n^{-2} \sum (X_i - m)(X_j - m) + 2(\bar{X} - m)^3$. Using the above relations and the value of $\mu_3(\bar{X})$ from Ex. 1, we get $E(M_3)$.

$M_4 = n^{-1} \sum \{(X_i - m)^4 - 4(X_i - m)^3(\bar{X} - m) + 6(X_i - m)^2(\bar{X} - m)^2 - 4(X_i - m)(\bar{X} - m)^3 + (\bar{X} - m)^4\} = n^{-1} \sum (X_i - m)^4 - 4n^{-2} \sum (X_i - m)^3 (X_j - m) + 6n^{-3} \sum (X_i - m)^2 (X_j - m)(X_k - m) - 3(\bar{X} - m)^4$, from which we get $E(M_4)$.

Using the values of $E(M_3)$, $E(M_4)$, $E(S^2)$ (from Ex. 2), we see that the mean values of $n^2 M_3/(n-1)(n-2)$ and $n^2[(n+1)M_4 - 3(n-1)S^4]/(n-1)(n-2)(n-3)$ are respectively κ_3 and κ_4 which lead to the second conclusion.

Now $\gamma_1 = \kappa_3/\sigma^3$, $\gamma_2 = \kappa_4/\sigma^4$. Using the above estimates of κ_3 and κ_4 and the unbiased estimate $s^2 = nS^2/(n-1)$ of σ^2 , we get the last result.

4. $E(X^k) = \alpha_k$, $\alpha_2(X^k) = E(X^{2k}) = \alpha_{2k}$ so that $\sigma(X^k) = \sqrt{\alpha_{2k} - \alpha_k^2}$. $X_1^k, X_2^k, \dots, X_n^k$ are mutually independent having the same mean α_k and standard deviation $\sqrt{\alpha_{2k} - \alpha_k^2}$ if α_{2k} exists, whence the result follows from the central limit theorem for equal components.

5. $\sigma(s^2) = \sigma(X^2)\sigma^2/(n-1) = \sigma^2 \sqrt{2(n-1)}/(n-1)$. Note that $\mu_4 = 3\sigma^4$.

6. (a) For a binomial (N, p) population $n\bar{X} = \sum X_i$ is binomial (nN, p) , since X_1, \dots, X_n are mutually independent each binomial (N, p) .

Answer: $x_i = i/n$ ($i = 0, 1, \dots, N$), $f_i = \binom{nN}{i} p^i (1-p)^{nN-i}$.

(b) For a Poisson- μ population, $n\bar{X} = \sum X_i$ is Poisson distributed with parameter $n\mu$. Answer: $x_i = i/n$ ($i = 0, 1, \dots$), $f_i = e^{-n\mu} (n\mu)^i / i!$

(c) Here $n\bar{X}$ is (nl) variate. Answer : $f(x) = n^{nl} e^{-nx} x^{nl-1} / \Gamma(nl)$
 $(0 < x < \infty)$.

7. $\text{cov}(\bar{X}, S^2) = E\{(\bar{X} - m)S^2\}$. Now $(\bar{X} - m)S^2 = n^{-2} \sum (X_i - m)^2 \times (X_j - m) - (\bar{X} - m)^3$. Using the above relations and $\mu_3(X)$ from Ex. 1, $\text{cov}(\bar{X}, S^2) = (n-1)\mu_3/n^2$.

14.7

1. $-1 - n/\log(x_1 x_2 \cdots x_n)$

2. For a sample of unit size $L = 2(a-x)/a^2$. $\frac{\partial \log L}{\partial a} = 0$ gives $a = 2x$. $E(\bar{X}) = \sigma/3$ so that $E(2\bar{X}) = 2\sigma/3 \neq \sigma$. Hence the estimate is biased.

3. $1/(1 + \bar{x})$

4. \bar{x} ; μ is the population mean, and hence the required conclusion.

5. $\sum (x_i - m)^2 / n$

6. \bar{x}

8. \bar{x}/l , $m = al$ which gives the conclusion.

9. For large n , \bar{X} is approximately normal $(\mu, \sqrt{\mu/n})$ so that $\sqrt{n/\mu}(\bar{X} - \mu)$ is approximately normal. Then $P(-u_\epsilon < \sqrt{n/\mu}(\bar{X} - \mu) < u_\epsilon) \simeq 1 - \epsilon$ or $P(A < \mu < B) \simeq 1 - \epsilon$, where A, B are the roots of the equation in μ : $n(\bar{X} - \mu)^2/\mu = u_\epsilon^2$.

10. $\sigma = a\sqrt{l}$. For large n , $\sqrt{n}(\bar{X} - al)/a\sqrt{l}$ is approximately normal $(0,1)$. $P(-u_\epsilon < \sqrt{n}(\bar{X} - al)/a\sqrt{l} < u_\epsilon) \simeq 1 - \epsilon$ or $P(A < a < B) \simeq 1 - \epsilon$, where A, B are the roots of the quadratic equation in a : $n(\bar{X} - al)^2 = l a^2 u_\epsilon^2$ etc.

11. By (14.5.1) : 14.6, 19.2

12. By (14.5.3) : 51.2, 62.8

13. By (14.5.3) m : 95% - (62.86, 65.99), length = 3.13, 98% - (62.45, 66.41), length = 3.96. By (14.5.6) σ : 95% - (1.23, 3.83), length = 2.60; 98% - (1.15, 4.44), length = 3.29

14. By (14.6.3) : (.51, .63)

15. By (14.6.3) : .18, .25

16. By Ex. 9 above : 4.59, 4.97

17. (a) By Ex.9 : (7.6, 8.8) (b) By (14.6.4) : (7.7, 8.7)
 18. By (14.6.4) : 22.42, 24.62
 19. $\bar{x}_1 - \bar{x}_2 \pm t_e \sqrt{(n_1 S_1^2 + n_2 S_2^2)/vn_1 n_2}$ where $v = n_1 + n_2 - 2$ and t_e is given by (14.5.4).
 20. v_1, v_2 - observed values of X_1, X_2 respectively. For large n_1, n_2 X_1, X_2 are approximately normal $(n_1 p_1, \sqrt{n_1 p_1 q_1})$ and $(n_2 p_2, \sqrt{n_2 p_2 q_2})$ respectively where $q_1 = 1 - p_1, q_2 = 1 - p_2$. Since $\hat{p}_1 = v_1/n_1, \hat{p}_2 = v_2/n_2$, we can take X_1, X_2 to be approximately normal $(n_1 p_1, \sqrt{v_1(n_1 - v_1)/n_1})$ and $(n_2 p_2, \sqrt{v_2(n_2 - v_2)/n_2})$ respectively, so that

$$\frac{(X_1/n_1 - X_2/n_2) - (p_1 - p_2)}{\sqrt{v_1(n_1 - v_1)/n_1^3 + v_2(n_2 - v_2)/n_2^3}}$$

is approximately standard normal etc.

15.5

- $r = 0.935$; $y - 8.65 = 0.879(x - 9.44)$, $x - 9.44 = 0.995(y - 8.65)$
- $r = 0.503$; $y - 71.4 = 0.496(x - 49.8)$, $x - 49.8 = 0.510(y - 71.4)$
- $y = 129.42 + 5.367x$ ($|r| = 0.950$)
 $y = 125.20 + 8.986x - 0.4524x^2$ ($R_y = 0.968$)
- (a) $y = -1.2249 + 1.3396x$ ($|r| = 0.9988$)
 (b) $y = 0.18488 + 1.0762x + 0.0080502x^2$ ($R_y = 0.9995$)
 (c) $y = 6.4109 + 0.039812x^2$ ($R_y = 0.9885$)
- $y = x^2 - 8.13x + 19.56$

16.9

- Statistic: $\chi^2 = \sum (x_i - m)^2 / \sigma_0^2$ whose sampling distribution under H_0 is χ^2 -distributed with n D.F. A left-tailed test for $H_1: \sigma < \sigma_0$ and a right-tailed test for $H_1: \sigma > \sigma_0$.
- Two-tailed χ^2 -test where χ^2 is the same as in Ex. 1.
- $H_0: m = 15.5$ against no alternative. By (16.6.1) $u = 1.20$. For $\alpha = .05$, by (16.6.2) C.R.: $|u| > 1.96$. H_0 is accepted.
- $H_0: m = .05$ against no alternative. By (16.6.3) $t = -4.819$. Two-tailed t -test. For $v = 11$, $P(|t| > 4.819) < .001$. Hence the value of t is highly significant, and we reject H_0 .

5. $H_0 : m = 0$ against no alternative. By (16.6.3) $t = 2$. Two-tailed t -test. For $\varepsilon = .05$, $v = 9$, by (16.6.4) C.R. : $|t| > 2.262$. Accept H_0 .

6. $H_0 : m = 65 = m_0$ against $H_1 : m < m_0$. By (16.6.3) $t = -0.870$. Left-tailed t -test. For $v = 7$, $P(t < -0.870) = .21$. Accept H_0 .

7. $H_0 : \sigma = 5.2$ against no alternative. By (16.6.5) $\chi^2 = 24.882$. Two-tailed χ^2 -test. For $\varepsilon = .05$, $v = 19$, by (16.4.9). C.R. : $0 < \chi^2 < 8.825$ and $\chi^2 > 33.096$. H_0 is accepted.

8. $H_0 : \sigma = 0.1 = \sigma_0$ against $H_1 : \sigma > \sigma_0$. By (16.4.5) $\chi^2 = 21.560$. Right-tailed χ^2 -test For $\varepsilon = .05$, $v = 10$, by (16.4.6) B.C.R. : $\chi^2 > 18.307$. Accept H_1 .

9. $H_0 : m_1 = m_2$ against no alternative. By (16.7.3) $u = 1.16$. For $\varepsilon = .01$, by (16.7.4) C.R. : $|u| > 2.58$. Accept H_0 .

10. $H_0 : m_1 = m_2$ against no alternative. By (16.7.5), $t = 1.50$. For $\varepsilon = .05$, $v = 18$, by (16.7.6.) C.R. : $|t| > 2.10$. H_0 is accepted.

11. $H_0 : m_1 = m_2$ against $H_1 : m_1 > m_2$ assuming $\sigma_1 = \sigma_2$. Right-tailed t -test ; $t = 4.384$ by (16.7.5). For $v = 18$, $P(t > 4.384) < .001$; value of t is highly significant so that H_0 is rejected. Since the second sample has a greater value of s^* , by interchanging the two populations $H_0 : \sigma_1 = \sigma_2$ against no alternative. Two-tailed F -test where $F = 1.27$ by (16.7.7). For $\varepsilon = .1$, $v_1 = v_2 = 9$, by (16.7.8) the right critical interval : $F > 3.19$. H_0 is confirmed.

12. $H_0 : m_1 = m_2$ assuming $\sigma_1 = \sigma_2$ against no alternative. By (16.7.5) $t = 0.749$. Two-tailed t -test. For $v = 13$, $P(|t| > 0.749) = .47$. Accept H_0 . $H_0 : \sigma_1 = \sigma_2$ against no alternative. Two-tailed F -test where $F = 2.31$ by (16.7.7). For $v_1 = 7$, $v_2 = 6$, $\varepsilon = .1$, by (16.7.8) C.R. : $F > 4.22$. H_0 is accepted.

13. $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 > \sigma_2$. By (16.7.7) $F = 4.93$. Right-tailed F -test. For $\varepsilon = .01$, $v_1 = 24$, $v_2 = 14$, by (16.7.8) C.R. : $F > 3.43$. H_0 is rejected.

14. $H_0 : \rho = 0$ against $H_1 : \rho < 0$. Left-tailed t -test where $t = -2.688$ by (16.8.1). For $v = 4998$, $P(t < -2.688) = .004$. Reject H_0 .

15. $H_0 : \rho = 0$ against no alternative. Two-tailed t -test where $t = 2.178$ by (16.8.1). For $v = 14$, $P(|t| > 2.178) = .048$; value of t simply significant so that we reject H_0 , but not very confidently.

17.7

1. $H_0 : p = 1/6$ against no alternative. By (17.1.1) $u = 4.19$. Two-tailed standard normal test. $P(|U| > 4.19) < .001$; value of u is highly significant. Reject H_0 .
2. $H_0 : p = 1/4$; no alternative. By (17.1.1) $u = 2.17$. For $\varepsilon = .05$ by (17.1.2), C.R. : $|u| > 1.96$. Reject H_0 .
3. $H_0 : p = \frac{1}{2} = p_0$ against $H_1 : p > p_0$. By (17.1.1) $u = 2.42$. Right-tailed standard normal test. $P(U > 2.42) = .008$; value of u is highly significant and so H_0 is rejected.
4. $H_0 : p = 0.1$; no alternative. By (17.1.1) $u = 2.17$. For $\varepsilon = .01$, by (17.1.2), C.R. : $|u| > 2.58$. Accept H_0 .
5. First coin : $H_0 : p_1 = 1/2$; no alternative. By (17.1.1) $u = -2.91$. Two-tailed standard normal test. $P(|U| > 2.91) = .004$; value of u is highly significant. Reject H_0 .
 Second coin : $H_0 : p_2 = 1/2$; no alternative. By (17.1.1) $u = -3.47$. Two-tailed standard normal test. $P(|U| > 3.47) < .001$ so that H_0 is rejected.
- $H_0 : p_1 = p_2$ against no alternative. By (17.2.1) $\hat{p} = 0.3295$, and by (17.2.2) $u = 0.20$. Two-tailed test. $P(|U| > 0.20) = .84$; value of u is not significant at all. Accept H_0 .
6. $H_0 : p_1 = p_2$ against no alternative. By (17.2.1) $p = 0.7832$, by (17.2.2) $u = 4.36$. Two-tailed test. $P(|U| > 4.36) < .001$. Reject H_0 .
7. $H_0 : \mu = 8$; no alternative. By (17.3.1) $u = 0.66$. Two-tailed test. $P(|U| > 0.66) = .51$; value of u is not significant at all. Accept H_0 .
8. $H_0 : \mu = 1/5$; no alternative. By (17.3.1) $u = -0.98$. Two-tailed test. $P(|U| > 0.98) = .33$; value of u is not significant. Accept H_0 .
9. $H_0 : p_k = 1/6$ ($k = 1, 2, \dots, 6$). Right-tailed χ^2 -test where $\chi^2 = 48.464$ by (17.5.1). For $v = 5$, $P(\chi^2 > 48.464) < .001$; value of χ^2 is highly significant. Reject H_0 .
10. $H_0 : p_1 = 9/16, p_2 = 3/16, p_3 = 1/4$. Right-tailed χ^2 -test where by (17.5.1) $\chi^2 = 3.70$. For $\varepsilon = .05, v = 2$, by (17.5.2) C.R. : $\chi^2 > 5.99$. Accept H_0 .

11. $H_0: p_1 = .1, p_2 = .5, p_3 = .4$. By (17.5.1) $\chi^2 = 1.296$. Right-tailed χ^2 -test. For $v = 2$, $P(\chi^2 > 1.296) = .53$; value of χ^2 is not significant at all. Accept H_0 .

12. (a) Combine the last 3 entries of the table. $\chi^2 = 20.09$. For $v = 3$, $P(\chi^2 > 20.09) < .001$; value of χ^2 is highly significant. Population distribution is not binomial (5, 1/6).

(b) Combine the last 3 entries of the table. p is replaced by $\hat{p} = 0.216$. $\chi^2 = 1.045$. For $v = 2$, $P(\chi^2 > 1.045) = .60$; value of χ^2 is not significant. Population distribution is binomial (5, p).

13. The first 2 and the last 3 entries of the table are combined together. μ is replaced by $\bar{x} = 8.20$. $\chi^2 = 2.924$. For $v = 7$, $P(\chi^2 > 2.924) = .89$. Population is Poissonian.

14. The first 2 and the last 2 entries of the table are combined together. m and σ are replaced by $\bar{x} = 0.79197$ and $S = 0.0117$ respectively. $\chi^2 = 6.746$. For $v = 6$, $P(\chi^2 > 6.746) = .35$. Population is normal.

18.7

1. From symmetry about the origin, X, Y have the same distribution, and let the density function of each be $f(x)$. Since X, Y are independent, $f(x, y) = f(x)f(y)$. Set $x = r \cos \theta, y = r \sin \theta$. The distribution of (X, Y) being symmetrical about the origin, $\frac{\partial}{\partial \theta} f(x, y) = 0$ which gives $f'(x)/xf(x) = f'(y)/yf(y) = k$, a constant, so that $f(x) = Ae^{kx^2/2}$. Now $\int_{-\infty}^{\infty} f(x) dx = 1$; for convergence of this integral $k < 0$, and setting $k = -1/\sigma^2$, we get $A = 1/\sqrt{2\pi} \sigma$. Hence etc.

2. $E(X) = a_1 m_1 + \dots + a_n m_n = m$ which shows that X can be regarded as a measured value of a quantity whose true value is m . The second part follows from (8.5.8) and (18.3.1) and the third from (18.5.2).

4. By (18.5.5) $W^{-1} \sum w_i e_i = \bar{x} - m$, and by (18.5.20) $v_i = e_i - W^{-1} \sum w_i e_i = -(w_1/W)e_1 - \dots + (1 - w_i/W)e_i - \dots - (w_n/W)e_n$. For the corresponding random variables $V_i = -(w_1/W)E_1 - \dots + (1 - w_i/W)E_i -$

$\dots - (w_n/W)E_n$. Since E_1, E_2, \dots, E_n are mutually independent, E_i being normal $(0, \sigma/\sqrt{w_i})$ ($i = 1, 2, \dots, n$), $\sigma^2(V_i) = \sum_{k \neq i} w_k \sigma^2/W^2 + (1 - w_i/W)^2 \times \sigma^2/w_i = (W - w_i)\sigma^2/Ww_i$. Hence etc.

5. $\bar{x} = 39.204$, $\sigma^+ = 0.224$, $Q^+ = 0.151$, $\sigma^+(\bar{X}) = 0.071$, $Q^+(\bar{X}) = 0.048$.
Confidence limits : 50% - 39.154, 39.254 ; 95% - 39.043, 39.365

6. $\bar{x} = 0.6793$, $Q_1^+ = 0.0044$, $Q_2^+ = Q_3^+ = Q_4^+ = 0.0036$, $Q_5^+ = 0.0031$, $Q_1^-(\bar{X}) = 0.0018$. Confidence limits : 0.6725, 0.6861

7. (a) $q_1^* = 1.59$, $q_2^* = 4.93$, $q_3^* = 1.47$

$m_1^* = 3.59$, $m_2^* = 14.38$, $m_3^* = 26.36$, $m_4^* = 10.11$

(b) $q_1^* = 1.56$, $q_2^* = 4.91$, $q_3^* = 1.47$

$m_1^* = 3.54$, $m_2^* = 14.31$, $m_3^* = 26.20$, $m_4^* = 10.03$

BIBLIOGRAPHY

- ARLEY, N. and BUCH, K. R. : Introduction to the theory of probability and statistics, Wiley, New York, 1950
- CHUNG, K. L. : Elementary probability theory with stochastic processes, Narosa, New Delhi, 1978
- CRAMER, H. : Random variables and probability distributions, Cambridge University Press, Second edition, 1961
- : Mathematical methods of statistics, Princeton University Press, 1958
- : The elements of probability theory and some of its applications, Wiley, New York, 1959
- FELLER, W. : An introduction to probability theory and its applications, Vol. I, Wiley Eastern, Third edition, 1978
- : An introduction to probability theory and its applications, Vol. II, Wiley Eastern, 1977
- FISHER, R. A. : Statistical methods for research workers, Oliver and Boyd, Edinburg, Eleventh edition, 1950
- and Yates, F. : Statistical tables, Oliver and Boyd, Second edition, 1943
- GNEDENKO, B. V. : The theory of probability, Mir, Moscow, 1969
- GOLDBERG, S. : Probability : An introduction, Prentice-Hall, New Jersey, 1960
- HOEL, P. G. : Introduction to mathematical statistics, Asia, Bombay, Second edition, 1961
- KENDALL, M. G. : The advanced theory of statistics, Vols. I-II, Griffin, London, 1945-48
- KENNEY, J. F. and KEEPING, E. S. : Mathematics of statistics, Part I, Van Nostrand, Third edition, 1954
- : Mathematics of statistics, Part II, Van Nostrand, Second edition, 1951

- KOLMOGOROV, A. N. : Foundations of the theory of probability, Chelsea, New York, 1950
- LEVY, H. and ROTH, L. : Elements of probability, Oxford University Press, 1936
- LINDGREN, B. W. and MCEL RATH, G. W. : Introduction to probability and statistics, Macmillan, New York, 1959
- MISES, R. V. : Probability, statistics and truth, William Hodge, London, 1939
- MOOD, A. M. and GRAYBILL, F. A. : Introduction to the theory of statistics, McGraw-Hill, New York, Second edition, 1963
- MUNROE, M. E. : Theory of probability, McGraw-Hill, New York, 1951
- NEYMAN, J. : First course in probability and statistics, Holt, New York, 1953
- PARZEN, E. : Modern probability theory and its applications, Wiley, New York, 1960
- ROZANOV, Y. A. : Introductory probability theory, Prentice-Hall, New Jersey, 1969
- SMART, W. M. : Combination of observations, Cambridge University Press, 1958
- USPENSKY, J. V. : Introduction to mathematical probability, McGraw-Hill, New York, 1937
- WEATHERBURN, C. E. : A first course in mathematical statistics, Cambridge University Press, Second edition, 1949
- WENTZEL, E. S. : Probability theory (First steps), Mir, 1982
- WHITTAKER, E. and ROBINSON, G. : The calculus of observations, Blackie, London, Fourth edition, 1944
- WILKS, S. S. : Elementary statistical analysis, Princeton University Press, New Jersey, 1958
- : Mathematical statistics, Princeton University Press, New Jersey, 1944

INDEX

Addition rule for mean values 153
 for probabilities 18, 23
 Additivity, complete 21
 Asymptotically normal 208
 Axioms 21

Bayes' theorem 35
 Bernoulli's theorem 201
 Buffon's needle problem 105

Cartesian product 46
 Characteristics 127, 222, 258
 Class frequency 220
 interval 220
 limits 220

mark or midpoint 220
 Classical definition 13
 Coefficient of excess 138, 224
 of kurtosis 138, 224
 of skewness 137, 224

Confidence coefficient 245
 interval 246
 limits 246

Convergence 'in probability' 198
 Correlation coefficient 155, 258
 ratio 182

Covariance 155, 258
 Critical region 274
 best 275

Cumulants 143, 239

Cumulative graph 218

Curve, density 81
 distribution 71

Curve fitting, parabolic 179, 264

Decile 148

Degrees of freedom 188, 192

Deviation, standard 135

Diagram, dot 258
 frequency 219
 probability 74
 scatter 258

Dispersion, measure of 134

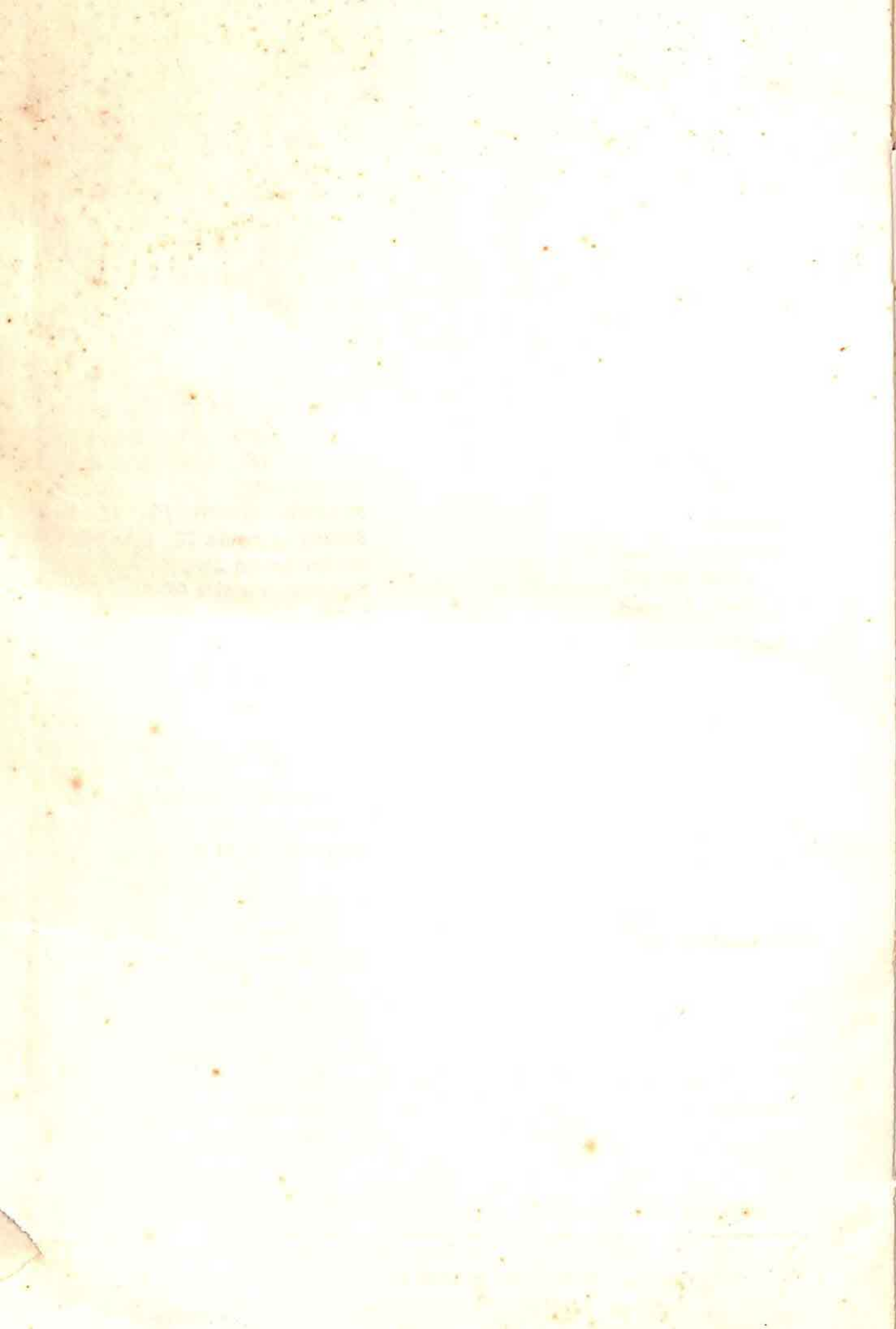
Distribution, binomial 75
 beta, of the first kind (β_1) 86
 „ of the second kind (β_2) 86
 Cauchy 85
 causal 74
 chi-square (χ^2) 188
 conditional 109
 continuous 80, 101
 discrete 72, 97
 F 194
 gamma (γ) 86
 Laplace 151
 log-normal 151
 marginal 95
 multinomial 307
 normal 84, 107
 of the sample 218
 Pascal 150
 Poisson 76
 rectangular 82, 104
 sampling 231
 Student's (t) 192
 uniform 82, 104

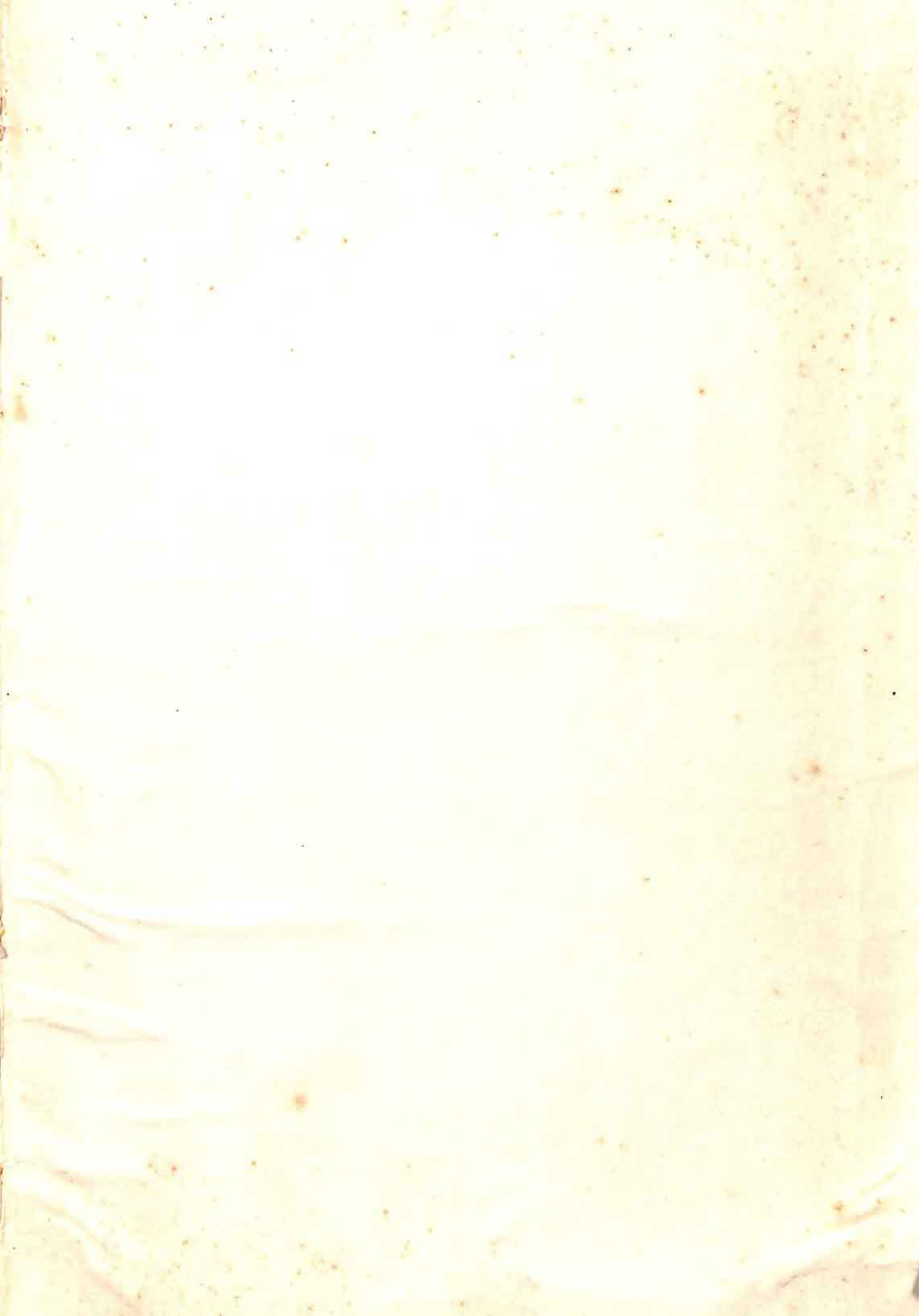
Equiprobability ellipses 108

Error 316
 accidental 316
 elementary 318
 experimental 1, 316
 mean square 321
 of observation 316
 probable 321
 random 316
 root mean square 321

- Error, systematic 316
 - two types of 274
- Estimate 233
 - consistent 233
 - unbiased 233
- Estimation, interval 245
 - point 245
- Event points 9
 - space 9
- Events 1, 3, 9
 - certain 4, 9
 - complementary 5, 9
 - compound 4
 - impossible 4, 9
 - mutually exclusive 4, 9
 - simple 4
- Expectation 127, 152
 - conditional 168, 169
- Frequency, absolute 16
 - conditional 18
 - definition 18
 - interpretation 21
 - ratio 16
 - relative 16
- Function, characteristic 140, 164
 - density 81, 101
 - distribution 69, 94
 - error 321
 - likelihood 241, 267
 - moment generating 139
 - step 72
- Gaussian law 320
- Grouping of data 220
- Histogram 220
- Hypothesis, alternative 272
 - composite 271
 - null 272
 - simple 271
 - statistical 271
- Independent events 39
 - random experiments 47
- Independent random variables 96
 - trials 48
- Kurtosis 138
- Law, binomial 50
 - multinomial 57
 - normal 317
 - of large numbers 202
- Least squares, principle of 173, 322, 330, 332, 335
- Liapounoff's condition 212
- Likelihood equations 242, 267
 - ratio 285
 - „ testing 283
- Limit theorem, central 209
 - deMoivre-Laplace 204
 - for characteristic functions 211
- Lindeberg's condition 209
- Location, measure of 139
- Markov chain 59
- Maximum likelihood estimates 242, 267
 - method 241
- Mean, 130, 222, 258
 - conditional 168, 169
 - weighted 327
- Mean value 127, 152
 - conditional 168, 169
- Median 145, 223
- Mode 147, 223
- Modulus of precision 321
- Moments 132, 155, 223, 258
 - absolute 132
 - central 132, 155, 223, 258
- Multiplication rule for mean values 162
 - for probabilities 19, 32
- Neyman-Pearson theorem 275
- Normal equations 174, 264, 322, 330, 333, 335

- Parametric point 272
 - space 272
- Peakedness 133
- Percentile 148
- Population 216
- Power of critical region 274
 - of test 274
- Probability, conditional 18, 31
 - differential 81, 102
 - element 81
 - mass 72
 - of transition 61
- Process, Poisson 77
 - stochastic 77
- Quantile 148
- Quartile deviation 148
 - lower 148, 223
 - upper 148, 223
- Random experiments 1, 2
 - observations 1
 - walk problem 167
- Random variable 66
 - normalised 136
 - standardised 136
- Range 224
 - semi-interquartile 148, 224
- Regression coefficient 175, 263
 - curves 169
 - „ least square 173
 - function 169
 - „ least square 174
 - linear 170
 - lines 175, 263
 - parabola 180
- Reproductive property 117, 165
- Residual 174, 264
- Residuals of the sample 264, 323
- Sample 216, 257
 - characteristics 222, 258
 - mean 222, 258
 - point 230
 - space 230
 - values 217
 - variance 222, 258
- Schwartz's inequality 185
- Semi-invariants 143
- Set 5
 - empty 6
 - null 6
- Significance level 275
 - tests of 273
- Skewness 137
- Statistic 231
- Statistical regularity 16
- Stirling's formula 27
- Student's ratio 239
- Stochastic variable 66
- Subset 5
- Sum polygon 218
- Tchebycheff's inequality 195
 - theorem 201
- Test 273
 - best 275
 - of goodness of fit 310
 - most powerful 275
- Trials, Bernoulli 49
 - „ infinite 58
 - Poisson 55
 - repeated 48
- Transformation of random variables 87, 113
- Transition matrix 61
- Variance 134, 222, 258
 - about regression function 171
 - conditional 168, 169
 - residual 178
- Variate 66
- Weight 326





519'9

GUP

219